Modeling Spatiotemporal Correlations between Video Saliency and Gaze Dynamics(Dissertation_全文)

AUTHOR(S):

Yonetani, Ryo

Modeling Spatiotemporal Correlations between Video Saliency and Gaze Dynamics

Ryo Yonetani

Abstract

This thesis addresses the problem of modeling the relationship between visual contents and human gaze data. This relationship serves as a glue to connect two research areas: visual content analyses and gaze behavior analyses. Namely, various techniques of visual content analyses can be leveraged for gaze behavior analyses thanks to the modeling of the relationship. We particularly cover a situation where humans are watching various videos taken in real environments. These videos are characterized by their time-varying scene structures formed by a variety of visual events such as object motions, texture variations, camera motions and scene changes. Facing such situations, we define the relationships named *spatiotemporal correlations* that refer to the following twofold relationships between video and gaze data: (1) event-level spatiotemporal gaps (e.g., temporal delays) between visual events in scene structures and corresponding reactions in gaze dynamics and (2) scene-level correlations between the scene structures and the gaze dynamics. Our goal is to develop a framework to describe the spatiotemporal correlations in a simple and efficient manner.

Our framework comprises the models of scene structures, gaze dynamics and spatiotemporal correlations. Since the scene structures appearing in our videos of interests are generally uncontrollable and unknown, we first need to extract visual events from given videos and model them to describe the scene structures. To this end, we particularly focus on the dynamic changes of salient regions that attract visual attention in videos and propose a series of models named the saliency dynamics models. The proposed models are capable of describing how various visual events influence gaze dynamics by using primitive spatiotemporal patterns of salient regions called saliency primitives. Namely, the scene structures are modeled by a set of saliency primitives. Since the primitives are achieved in a data-driven fashion, obtained scene structures are efficient for given videos. Besides, we describe gaze dynamics with a sequence of primitive patterns as well, which we refer to as gaze primitives. Then, the spatiotemporal correlations can

be simply modeled as the relationships among primitives. Specifically, the event-level spatiotemporal gaps can be described with the spatiotemporal distances defined in a pair of saliency and gaze primitives. In addition, the scene-level correlations can be modeled as the combinations of saliency primitive sets (i.e., the scene structures) and gaze primitives in a certain temporal interval.

The effectiveness of our framework is assessed by describing spatiotemporal correlations and evaluating them via several practical gaze behavior analyses in real environments. First, we address the special situation where scene structures are constant and visual events are given so that we just need to focus on spatiotemporal correlations. Specifically, we aim to capture temporal synchronizations between visual events being focused on and corresponding gaze reactions when observers are browsing a content. Within our framework, these synchronizations can be described with the temporal distances between the onset times of saliency and gaze primitives, which is one aspect of the event-level spatiotemporal gaps. We confirmed the effectiveness of this description via attentional target identification tasks.

In the next step, we test our framework with intentionally-designed videos containing time-varying scene structures with various visual events, and standing for a more practical situation. By adopting saliency dynamics models, we identify the types of scene structures that vary over time and investigate how different scene-level correlations can appear depending on those types. Particularly, we describe the scene-level correlations based on the types of gaze primitives or those of saliency primitives and statistically learn them for each type of scene structures. The learned results made a great contribution to attentive state estimation tasks when watching TV commercial films.

Based on the above two studies that address event-level spatiotemporal gaps and scene-level correlations separately, we finally focus on overall spatiotemporal correlations. Specifically, this part aims to describe how spatiotemporal gaps are influenced by scene-level correlations. We first introduce a model of gap structures that jointly describe the gaps and scene structures with help from saliency dynamics models. Then, the modeled gap structures are statistically learned with respect to each type of gaze primitives to involve the scene-level correlations between scene structures and gaze dynamics. The experiments of gaze-point prediction from videos revealed that the learned results of spatiotemporal correlations were able to explain gaze behavior well when observers are watching various categories of videos including unedited natural ones.

Acknowledgment

I first wish to express my deepest gratitude to my supervisor, Professor Takashi Matsuyama. He has hospitably guided me since I was an undergraduate student. His insightful and constructive comments covered not only the main concepts of this thesis but also its technical details. And furthermore, he has always shown me a great vision for a way of life as a researcher.

I also wish to express sincere appreciation to my thesis committee, Professor Toshio Inui and Professor Shin Ishii, for taking time to review my thesis and giving many helpful comments.

Dr. Hiroaki Kawashima at Matsuyama Laboratory and Dr. Takatsugu Hirayama at Nagoya University have always supported me wherever they were. This thesis cannot be completed without their generous help. I am also thankful for Dr. Shohei Nobuhara, Dr. Tony Tung, Dr. Takekazu Kato, Dr. Takeshi Takai and Dr. Xuefeng Liang for their daily assistance. Special thanks go to our secretaries, Ms. Shiho Kimura, Ms. Aya Inoue, Ms. Mayumi Izumi, Ms. Yukimi Ishida, Ms. Asako Yoshimura and Ms. Kazuyo Hashimoto, and to all the members of Matsuyama Laboratory, especially but not limited to Ms. Erina Ishikawa, Mr. Kento Tamura, Mr. Kei Shimonishi, Mr. Pablo Roman Humanes, Mr. Qun Shi, Ms. Cuicui Zhang and Mr. Yoshinori Tarumoto. I really enjoyed my daily life in the laboratory.

I have been blessed with many like-minded colleagues, Mr. Hiroshi Kajino at The University of Tokyo, Mr. Koichiro Yoshino, Mr. Takuma Otsuka, Ms. Shoko Fujita-Tarumoto and Mr. Keisuke Otaki at Kyoto University.

Finally, I wish to express my greatest love and gratitude to my wife Risa.

Contents

Chapter 1

Introduction

1.1 Motivation

We humans are surrounded by a vast amount of display systems in our daily life. These systems provide visual contents that involve a variety of visual events such as scene changes in movies, human actions in surveillance videos and camera motions in first-person-view videos. Facing such contents, we direct our eyes to them and try to get information by design. Alternatively, eyes are sometimes directed to the contents unconsciously when eye-catching events happen such as sudden pop-ups of logos in commercial films.

Researchers have long studied visual contents and humans mainly in the fields of computer vision, human computer interaction (HCI), multimedia, visual psychology and neuroscience. As will be reviewed in the next section, their interests loosely fall into two issues: analyzing visual contents themselves (**visual content analyses**) and analyzing how humans act when they face the contents (**human behavior analyses**). Above all, an eye movement is one of the most important modalities that strongly reflect both internal mental states of humans and external visual events in the contents. Gaze behavior analyses and their applications in real environments are indeed one of the recent trends in numerous research fields: for example, for measuring gaze-based social interactions [Park et al., 2012, Fathi et al., 2012], mental state estimation from gaze [Brandherm et al., 2007, Simola et al., 2008, Nakano and Ishii, 2010, Hirayama et al., 2010, Eivazi and Bednarik, 2011, Bednarik et al., 2012], proficiency assessments [Eivazi et al., 2012], detection of developmental disorders [Tseng et al., 2013], gaze-based content designs [Simonin et al., 2005] and gaze-based recommender systems [Yoshitaka et al., 2007].

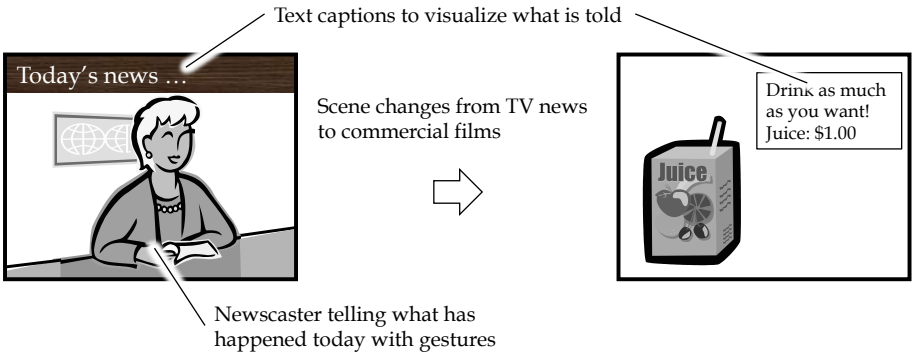

Figure 1.1: Visual events in videos.

Based on the two research tides, this thesis aims to develop a framework to describe the relationships formed by visual contents and gaze data which occur when human observers are watching the contents. Specifically, we extract and model visual events and their influences upon observers' gazes via visual content analyses in the proposed framework. In addition, the effectiveness of the framework is assessed via practical gaze behavior analyses in real environments.

1.2 Background

This section briefly reviews selected milestone studies on visual content analyses and gaze behavior analyses. Within the studies on the visual content analyses, we mainly review what can serve as visual events and how they have been addressed in the field of computer vision. With regard to the gaze behavior analyses, we review human vision studies on the relationships between eye movements and visual inputs as well as several practical techniques of mental state estimation as an example of the gaze behavior analyses.

1.2.1 Visual Content Analyses

Imagine that we are watching TV news like Figure 1.1. In the video, a newscaster tells us what has happened today with gestures. At the same time, the video can display text captions and photos to visualize what was told. In addition, the video sometimes contains scene changes from news to commercial films.

The preceding example contains various visual events where the events can be categorized based on several aspects. Figure 1.2 overviews the considerable aspects of visual events and examples of the events categorized by the aspects.

		Local aspects	Global aspects
Physical aspects	Static aspects	- Positions of objects - Shapes of objects - Textures in spatial patches - Spatial saliency	- Object layouts
	Dynamic aspects	- Appearances/disappearances of objects - Translations of objects - Deformations of objects - Texture variations in spatiotemporal patches - Spatiotemporal saliency	- Scene changes - Camera motions
Semantic aspects	Static aspects	- Object categories - Texture categories	- Scene categories
	Dynamic aspects	- Action categories	- Video categories

Figure 1.2: Aspects and examples of visual events.

As for Figure 1.1, spatiotemporal variations such as motions and deformations of objects resulted from gestures belong to physical aspects while the category of gestures (e.g., handwaving) are semantic ones. In addition, positions of newscasters are local and static aspects and scene changes are global and dynamic aspects. These aspects and several relevant studies so far can be summarized as follows.

Physical aspects and semantic aspects

Physical aspects refer to spatial or spatiotemporal signals resulted from visual events while semantic ones are generally a label given to the physical signals to explain their meaning. Whichever aspects we are interested in, the analyses begin with the extraction of physical-level (low-level) features such as the scale-invariant feature transform [Lowe, 1999], histograms of oriented gradients [Dalal and Triggs, 2005], optical flows (e.g., [Brox et al., 2004]), histograms of oriented optical flows [Chaudhry et al., 2009] and space-time interest points [Laptev, 2005]. Recognition of semantic aspects from these features will be summarized by associating them with other aspects in the following subsections.

On physical aspects, Review [Yilmaz et al., 2006] summarizes the representation of objects for object tracking techniques. In addition, modeling of shapes (object contours) is an important topic in relevant fields and several key contributions have been proposed so far: for example, Snakes [Kass et al., 1988], Active Appearance Models [Cootes et al., 2001] and Level Sets [Cremers, 2006]. The modeling of saliency (i.e., how a certain input differs compared to their surrounds) is also addressed by many studies and will be particularly reviewed in Chapter 2.

Static aspects and dynamic aspects

Visual events appear in the static form of objects or textures at each frame and they vary over time to generate dynamic spatiotemporal patterns. Representations of the spatiotemporal patterns can be roughly categorized into direct representations and model-based representations. The direct representations respectively utilize vectors and matrices (multidimensional vector sequences) to describe univariate and multivariate spatiotemporal patterns. Since these representations bring diversity as the size of patterns become larger, summarization techniques are sometimes adopted such as piecewise aggregate approximation [Yi and Faloutsos, 2000] and symbolic aggregate approximation [Keogh et al., 2005] (see Review [Ratanamahatana et al., 2005]). On the other hand, model-based representations adopt a parametric model such as linear dynamical systems (LDS), hidden Markov models (HMM) and segment models to summarize the patterns (see Review [Ostendorf et al., 1996]). For example, [Doretto et al., 2003] has proposed the dynamic texture model to describe dynamic changes of textures in a certain spatiotemporal patch based on the LDS. In addition, [Chan and Vasconcelos, 2008] and [Ravichandran et al., 2009] have introduced the mixtures of dynamical systems to represent more complex dynamic textures. Human motions often consist of the switches of several primitive patterns (e.g., swinging hands consists of the switches of left-to-right motions and right-to-left motions). Thus, several studies have introduced switching linear dynamical systems (SLDS) to model such complicated human motions [Bregler, 1997, North et al., 2000, Li et al., 2002].

As for semantic aspects, static semantics include the categories of objects, textures and scenes while those of actions serve as dynamic semantics. Detection and recognition of such categories is one of the central issues in the field of computer vision and indeed numerous techniques have been proposed so far: detection techniques of faces [Viola and Jones, 2001], humans [Dalal and Triggs, 2005], generic objects [Lowe, 1999, Felzenszwalb et al., 2010], recognition techniques of scenes [Fei-Fei and Perona, 2005, Li et al., 2011], videos (Review [Brezeale and Cook, 2008]), actions (Review [Poppe, 2010]), etc.

Local aspects and global aspects

Visual events can take place both locally and globally. Several global aspects are formed by a set of local aspects, which is a central assumption in con-

tent analysis techniques based on the bag of visual words (BoVW) such as the scene recognition proposed in [Fei-Fei and Perona, 2005]. In addition, spatiotemporal layouts of local events can be also a crucial clue for scene recognition [Li et al., 2011, Harada et al., 2011, Sadeghi and Farhadi, 2011]. The BoVW contributes to the recognition of local aspects as well, such as object recognition [Sivic and Zisserman, 2003] and action recognition [Dollar et al., 2005]).

Several global events occur independently of local events: camera motions and scene changes, for example. Camera motions, such as panning, tilting and zooming, can be generally obtained by discriminating global motions from optical flows (e.g., [Rath and Makur, 1999, Chan and Vasconcelos, 2008, Zhang et al., 2013]). In addition, scene changes can be detected via video shot segmentation (see Review [Cotsaces et al., 2006]).

1.2.2 Gaze Behavior Analyses

Terminology

Before starting the review of gaze behavior analyses, we first need to specify the meaning of several technical terms. The following terminology is crucial since several terms are indeed used in a different way between computer vision and human vision fields (e.g., eye movements and saliency).

The relationships between eye movements and visual inputs are often described in a framework of visual attention mechanisms. We here draw some important concepts from the taxonomy presented by [Tsotsos, 2011].

Overt/covert attention is an action to select a stimulus from visual scenes. Overt attention is particularly an action to capture the stimulus into the fovea with eye movements, while covert one does not accompany the eye movements. When literature poses a confrontation like "visual attention and eye movements" [Hoffman, 1998], the visual attention often specifies the covert one. However, we will use the term visual attention (or just attention) only when we need not specify its type.

Eye movement is one of observed characteristics resulted from the overt attention[(i)]. There are various types of eye movements such as **fixations** (maintaining points of gaze at a certain stimuli), **pursuits** (tracking a stimulus

[(i)] As presented in [Tsotsos, 2011], not only eye movements but head and body motions can be interpreted as observed characteristics. In this thesis, however, we particularly focus on the eye movements for simplicity.

in motion by a smooth movement) and **saccades** (rapidly shifting points of gaze). Although they are originally the rotations of eye-balls, sometimes they refer to shifts of gaze points on a screen in computer vision and HCI fields. In this thesis, we follow the original meaning and refer to the eye-ball rotations as the eye movements.

Exogenous influence is attentional signals that come from external visual stimuli. The attention resulted from the exogenous influences is referred to as **exogenous attention**.

Saliency is one of the important properties of exogenous visual stimuli that attract visual attention in a bottom-up manner. The degree of saliency is originally given by the contrast of stimuli between a certain point and its surround. However, the saliency has been also used to describe something conspicuous or important recently (see Reviews [Kimura et al., 2013, Eckstein, 2011]). In this study, we follow the original definition of saliency.

Endogenous influence is attentional signals that come other than external stimuli (e.g., knowledge, tasks and preference).

Inhibition of return is a mechanism to give a bias against returning visual attention to the locations (or objects) previously attended. Note that we do not particularly deal with the inhibition of return in this study since we assume the situation where visual inputs can vary over time.

Although the preceding concepts explain what can happen and affect human insides, what is actually observed from gaze tracking systems is the only sequences of gaze points on a screen, and this is our focus. We therefore introduce several terms as to the observed data.

Gaze data consist of gaze points on a screen provided via gaze tracking techniques (see Review [Morimoto and Mimica, 2005]).

Gaze dynamics refer to a physical spatiotemporal pattern consisting of a sequential shift of gaze points. Since they reflect vertical and horizontal rotations of eyeballs, we sometimes classify them based on biological definitions of eye movements such as fixations, pursuits and saccades.

Gaze behavior describes the overall behavior regarding gazes. It consists of conscious gaze actions with particular mental states (e.g., interests, intentions

and attentive states) as well as unconscious gaze dynamics. Thus, the gaze behavior analyses in this study include not only recognition tasks of gaze actions (e.g., mental state estimation) but also prediction tasks of conscious and unconscious gaze dynamics (e.g., gaze point prediction).

Spatiotemporal gaps between visual inputs and gaze reactions

Given a visual input like a certain visual event, human vision studies analyze how humans react to the input (e.g., the visual search reviewed in Chapter 2). Particularly when the visual input takes the form of videos, we sometime observe a kind of spatiotemporal gaps between a pair of inputs and reactions, which is a crucial phenomenon in this study. For example, we sometimes fail to direct our eyes to salient objects captured in peripheral vision when the objects have already moved or gone out of the frame before the shift of gaze. Besides, when focusing on salient objects in fast motion, the eyes are sometimes directed to different locations from object regions since it is hard to keep our eyes on the regions.

If we assume covert attention is still oriented to objects of interests while eyes are directed to other locations in the situations presented above, the spatiotemporal gaps can be regarded as the relationships between "visual attention and eye movements" [Hoffman, 1998], which has been well studied from the early 20th century [Kowler, 2011]. Generally, eye movements are believed to require preceding shifts of covert attention in several cases. For example, preview effects in reading are the phenomena that humans fixate a word in a sentence while attending about-to-fixated word in their periphery [Rayner, 1975]. In addition, the disassociation of covert attention and saccadic eye movements, that is, a situation where humans move their eyes and their spatial attention to different locations, is discussed in [Hoffman, 1998]. With regard to temporal gaps, we can predict a trajectory of object motions, and attend the destination before the object arrives. Smooth pursuit can be indeed initiated before the beginning of object motion, which is referred to as anticipatory smooth eye movements [Kowler, 2011]. In addition, [Mathot and Theeuwes, 2010] revealed that there existed a predictive remapping mechanism in visual attention that played an important role to predict where target would appear next. As a study on reaction delays, [Rashbass, 1961] has investigated a saccadic response toward an object with sudden motions. It has revealed that humans required saccades with reaction delays before pursuits if they were attending an object in motion. [Joiner and Shelhamer, 2006] has examined the similarity in the reaction delays between pursuits and saccades.

Figure 1.3: Scene structures consisting of multiple visual events. Parts of the images in this figure are contained in the dataset provided by [Mahadevan and Vasconcelos, 2010].

Finally, several studies in the field of visual psychology have addressed eye movements when watching a variety of videos including movies [Goldstein et al., 2007], animations [Munn et al., 2008] and dynamic natural scenes [Dorr et al., 2010]. They have particularly focused on the similarities of eye movements among subjects.

Mental state estimation

Eyes are a window into the mind — eye movements reflect various cognitive states. The first milestone is Yarbus's study (see [Yarbus, 1967]) that has revealed many important findings including the similarity of eye movements toward the same images or the dissimilarity depending on given tasks.

The findings above indicate the possibility to bring research schemes from tests of statistical significances in gaze behavior to classification problems of mental states from gaze data via statistical pattern recognition and machine learning. Indeed, several studies have proposed a method to estimate various states including interests [Hirayama et al., 2010], intentions and engagement levels in human-computer/agent interaction [Nakano and Ishii, 2010, Bednarik et al., 2012], intentions in problem-solving tasks [Eivazi and Bednarik, 2011] and those in search tasks [Simola et al., 2008, Ishikawa et al., 2012]. The basic approach of these method is to adopt a supervised learning framework with features extracted from gaze data and labels of mental states given in a top-down manner. As listed in [Jacob and Karn, 2003], numerous kinds of gaze features have been proposed: fixation durations, the number of fixations, scan path directions, saccade frequencies and lengths, for example. In addition, mental states are generally modeled as one of several discrete states.

1.3 Contributions

1.3.1 Preliminary Assumptions, Aims and Central Issues

Let us assume a situation where a single human observer is watching various videos taken in real environments, such as TV news, commercial films, surveillance videos and dynamic interfaces placed in airports, shopping malls and so on. In addition, we assume that gaze data from the observer watching the videos are obtained by gaze tracking systems.

As reviewed so far, our videos of interests contain various kinds of visual events that can attract our eyes. In this thesis, we take particular note of the dynamic aspects of visual events and aim to describe how dynamic changes in the events influence gaze dynamics. Moreover, we here introduce the term *scene structures* that refer to overall properties of video scenes consisting of various visual events. For example, Figure 1.3 depicts a scene structure consisting of an appearance and translations of objects (the pedestrian and car in the figure). We first aim to propose a novel approach to the modeling of visual events and scene structures so that we can deal with their influences on the gaze dynamics.

Given a modeled scene structure, we develop a framework to describe the relationships between video and gaze data, which comprises the models of gaze dynamics and the relationships themselves as well as those of scene structures. The effectiveness of our framework, in other words, how the framework can describe actual situations and contribute to practical applications, are assessed by describing the relationships and evaluating them via practical gaze behavior analyses in real environments.

Towards the above aims, we address the following two issues.

Issue 1: Handling diverse visual events in videos. As reviewed in Section 1.2.1, the recognition and understanding of visual events are still an active research topic in the field of computer vision. The main difficulty in the analyses lies in their diverse physics and semantics. The videos taken in real environments can display numerous types of dynamic changes in the form of spatiotemporal patterns, and at the same time, those changes are given a variety of semantic category labels. Moreover, it is generally uncontrollable and unknown that when, where and what kinds of visual events take place in the videos. These natures of visual events also bring difficulties when analyzing the relationships between video and gaze data in this study.

Issue 2: Considering time-varying scene structures. Due to the diverse visual events posed above, scene structures can vary over time. Then, we need to consider the following twofold relationships: (1) visual events in a scene structure influence gaze reactions to the events, and (2) scene structures influence overall gaze dynamics being observed. For example, (1) objects in motion can cause a reaction delay in pursuit gaze reactions, but (2) it depends on the types of scene structures (e.g., if they contain moving objects or not) that if gaze dynamics originally contains the pursuits. It is a different situation from traditional human vision and HCI studies that aim to clarify gaze behavior under controlled situations. They generally assume constant or a limited type of scene structures and visual events to observe specific gaze dynamics. In conclusion, a novel framework is required to describe the relationships when dealing with the time-varying scene structures.

1.3.2 Saliency Dynamics Models to Describe Scene Structures

As for Issue 1, our first contribution is to propose a model named *saliency dynamics models* that describes dynamic aspects of visual events. The basic idea is to particularly focus on the influence of visual events upon gaze dynamics instead of recognizing their semantic aspects. This idea is aimed at avoiding semantic diversity of visual events while preserving the essence when describing the relationships between video and gaze data. To this end, we aim to leverage the dynamic changes of saliency in videos for event characterizations. Specifically, we extract spatiotemporal patterns of salient regions from videos, which we refer to as saliency dynamics (Figure 1.4 (1)).

To describe the saliency dynamics extracted above, the proposed models introduce a primitive spatiotemporal pattern of salient regions referred to as *saliency primitives* (Figure 1.4 (2)). The saliency primitives serve as a unit to describe the saliency of various local and global dynamic events such as the ones listed in Figure 1.2. Namely, they indicate how much visual events attract our attention while sacrificing why they attract the attention explained by semantic aspects. In addition, a set of the primitives can characterize overall scene structures and thus they can contribute to the description of time-varying scene structures posed in Issue 2 (Figure 1.4 (3)). By achieving saliency primitives from videos in a data-driven manner, we can describe scene structures efficiently for given videos.

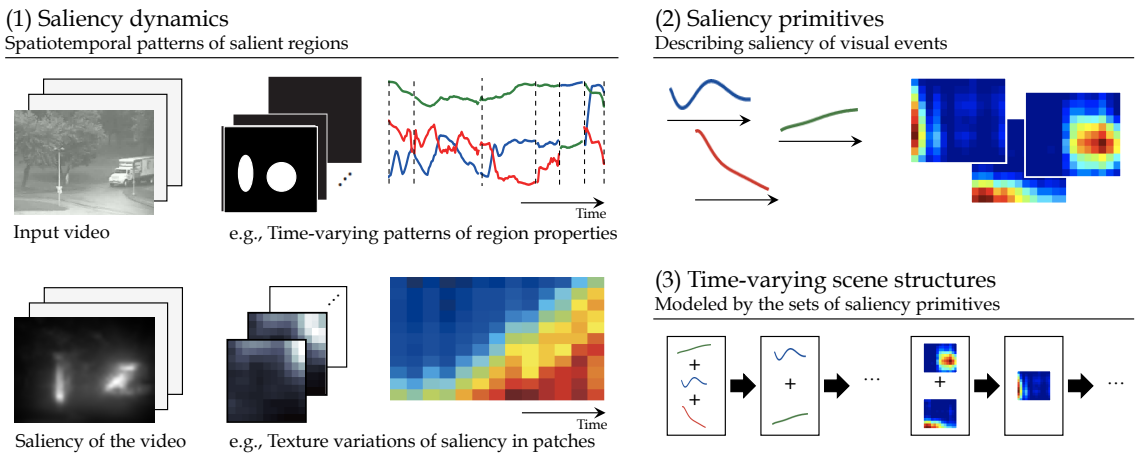

Figure 1.4: Overview of saliency dynamics models. Parts of the images in this figure are contained in the dataset provided by [Mahadevan and Vasconcelos, 2010].

1.3.3 A Framework for Spatiotemporal Correlations

The second contribution is the development of a framework to describe the relationships between video and gaze data. While scene structures of videos can be handled by the saliency dynamics models, we need models of gaze dynamics and the relationships as well, where the relationships involve the twofold characteristics presented in Issue 2. As for gaze dynamics, we first introduce the following generative process of gaze behavior based on the terms introduced in Section 1.2.2 as the basis of modeling.

Generative process of gaze behavior (see also Figure 1.5):

1. Videos containing various visual events are provided to human observers as visual inputs.

2. Scene structures modeled by the sets of saliency primitives influence attentional selections in an exogenous manner. The primitives characterize the saliency of visual events and they serve as attentional targets.

3. At the same time, mental states also influence the selection mechanism in an endogenous manner. For simplicity, we model them as one of several discrete states such as high/low levels of attentiveness.

4. Observers select one of saliency primitives as attentional targets and perform overt attention, where the selections are influenced by both scene

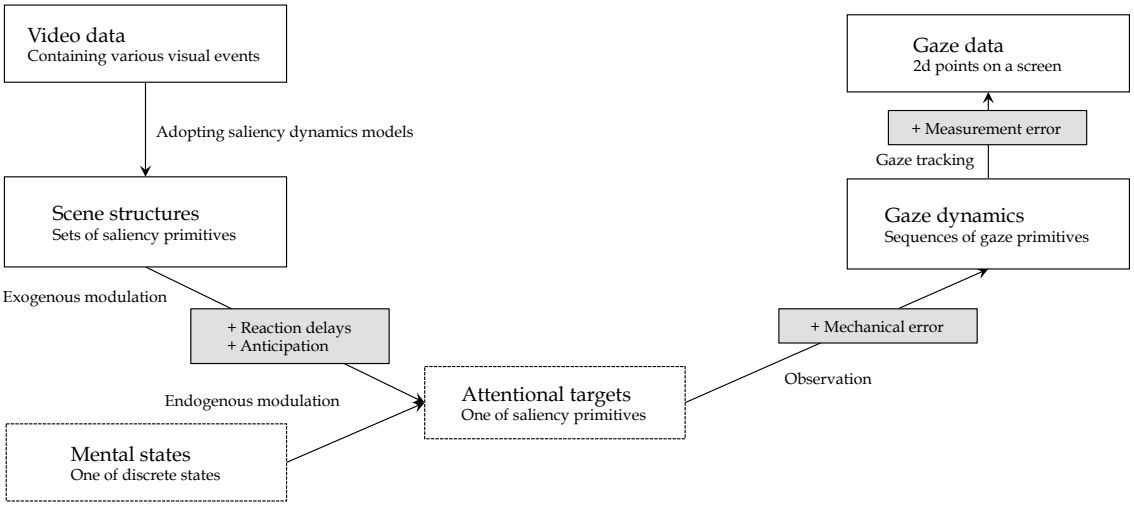

Figure 1.5: Generative process of gaze behavior.

structures and mental states. The selection can involve reaction delays and anticipation to the targets when observed saliency primitives contain specific dynamics such as sudden motions and appearances.

5. The overt attention accompanies eye movements to capture the target into the fovea, and the eye movements provide gaze dynamics. The gaze dynamics can contain some errors resulted from the mechanical systems of eye-ball rotations.

6. The gaze dynamics can be observed and stored as gaze data via gaze tracking. The observed data can involve measurement errors depending on a gaze tracking accuracy.

Based on the aforementioned process, gaze dynamics can be decomposed into primitive patterns since observers' gaze dynamics are assumed to be resulted from input saliency primitives. We thus model the gaze dynamics as sequences of the primitive patterns, which we refer to as *gaze primitives*. Along with the modeling of saliency dynamics, we can describe the gaze dynamics simply and efficiently by achieving gaze primitives in a data-driven manner.

Thanks to the primitive-based descriptions of scene structures and gaze dynamics presented so far, we can model the relationships between video and gaze data simply as the relationships among primitives. Specifically, we now introduce the special term *spatiotemporal correlations* to describe the twofold relationships posed in Issue 2. The spatiotemporal correlations consist of *event-level spatiotemporal gaps* and *scene-level correlations* of the following characteristics:

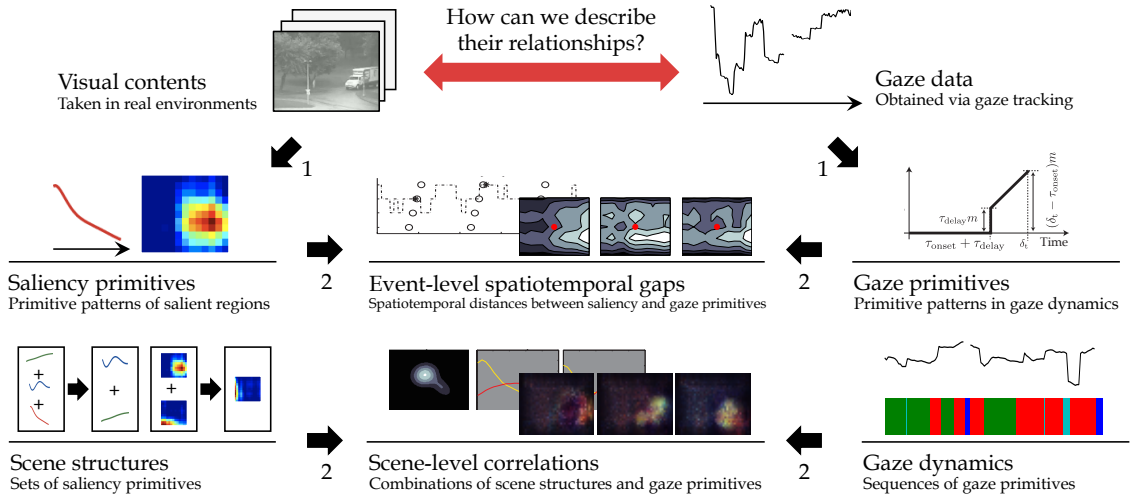

Figure 1.6: Framework describing the spatiotemporal correlations between video and gaze data. Parts of the images in this figure are contained in the dataset provided by [Mahadevan and Vasconcelos, 2010].

Event-level spatiotemporal gaps are temporal or spatiotemporal distances defined in a pair of saliency and gaze primitives, which aim to explain the influences from a single visual event to the corresponding gaze reaction. As shown in the generative process of gaze behavior, there are various factors that bring spatiotemporal gaps between primitives, such as reaction delays and anticipation in when reacting to a certain visual event.

Scene-level correlations are the combinations of modeled scene structures, i.e., sets of saliency primitives, and (possibly sequences of) gaze primitives in a certain temporal interval. Dynamic changes in the types of these correlations over time can explain the influences from time-varying scene structures to the gaze dynamics posed in Issue 2.

Consequently, the proposed framework comprising the models of scene structures, gaze dynamics and spatiotemporal correlations is summarized as Figure 1.6. The framework first receives video and gaze data to extract primitives (Arrows 1 in Figure 1.6) and exploit them for describing their event-level spatiotemporal gaps and scene-level correlations (Arrows 2 in the figure).

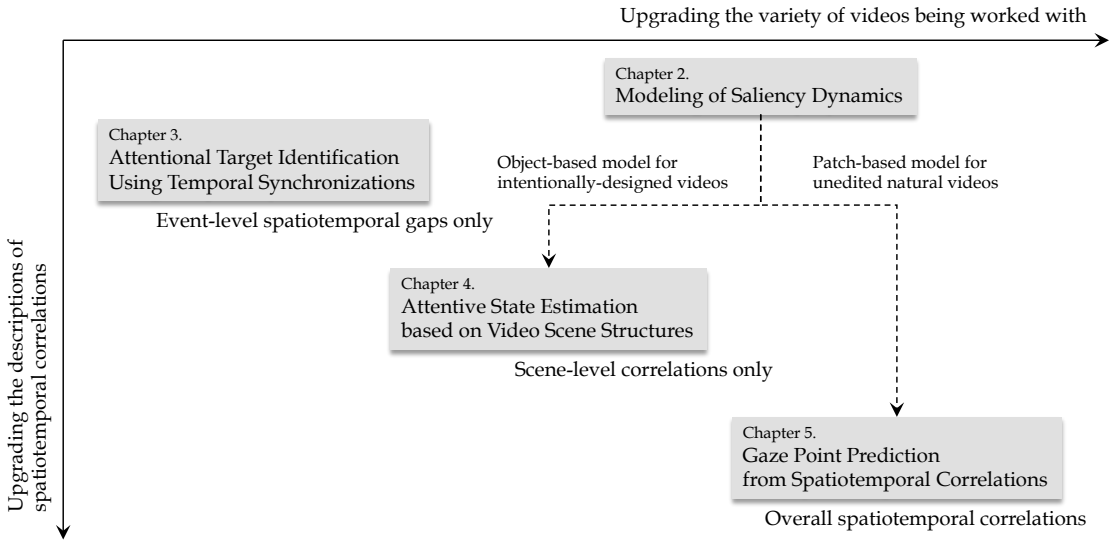

Figure 1.7: Roadmap of this study.

1.4 Structure of the Thesis

As an introduction to the following chapters, we briefly overview the positions and contributions of the component studies in this thesis. Each chapter can be mapped onto the roadmap shown in Figure 1.7. We first propose saliency dynamics models in Chapter 2, and gradually upgrade a variety of videos being worked with in the subsequent chapters. The effectiveness of our framework is assessed by describing spatiotemporal correlations and evaluating them via several practical gaze behavior analyses in real environments. In this context, Chapters 3, 4 and 5 individually focus on different aspects of the spatiotemporal correlations in incremental steps with different tasks of gaze behavior analyses.

Chapter 2. Modeling of Saliency Dynamics

This chapter first introduces a series of saliency dynamics models to describe visual events and time-varying scene structures. Physical dynamic aspects of various visual events presented in Figure 1.2 are reflected in the variations of saliency maps. The saliency primitives, primitive spatiotemporal patterns of salient regions, serve as a descriptor of those variations. We discuss several options for the modeling of saliency primitives to appropriately deal with visual events, and propose two practical saliency dynamics models named *object-based saliency dynamics model* and *patch-based saliency dynamics model*.

The object-based saliency dynamics model (OSDM) is aimed at handling visual events and scene structures in intentionally-designed videos such as TV commercial films. The primitive of the OSDM parametrically describes time-varying patterns of salient region properties to deal with visual events caused by distinct objects (e.g., object translations and deformations). In addition, the OSDM involves a scene segmentation mechanism so that it can also describe scene change events by the switches of scene structures.

The patch-based saliency dynamics model (PSDM), on the other hand, aims to cope with unedited natural videos like surveillance videos. The PSDM directly describes texture variations of saliency appearing in a certain spatiotemporal patch as saliency dynamics patterns. While the PSDM allows us to deal with a more variety of visual events including texture variations, it requires an efficient description of the dynamics patterns in the patch. We thus introduce a codebook of saliency primitives describing localized parts of the dynamics patterns, and aim to learn the codebook from videos.

Chapter 3. Attentional Target Identification Using Temporal Synchronizations

This chapter first evaluates the proposed framework under the special situation where we just need to focus on spatiotemporal correlations. Specifically, we manually design a constant scene structure of dynamic contents consisting objects generating a given type of visual events, and we analyze gaze behavior when human observers freely watch the contents.

We particularly aim to describe event-level spatiotemporal gaps that appear as temporal synchronizations between a pair of visual events and corresponding gaze reactions when observers are focusing on objects with the events. Within our framework, these synchronizations can be described with the temporal distances between the onset times of saliency primitives representing dynamic changes of given visual events and those of corresponding gaze primitives detected from gaze data, which are one aspect of the event-level spatiotemporal gaps. We leverage the temporal synchronizations for the task of attentional target identification that estimates which objects in contents are focused on.

Chapter 4. Attentive State Estimation based on Video Scene Structures

While the designed contents in the previous chapter contain constant scene structures with a given type of visual events, videos taken in real environments involve

time-varying scene structures due to various visual events. This chapter focuses on intentionally-designed videos with the events of frequent scene changes and various object motions. We describe them by the OSDM introduced in Chapter 2.

Unlike the previous chapter, this part mainly addresses the description of scene-level correlations in the spatiotemporal correlations. Specifically, we investigate how the scene-level correlations can be characterized differently depending on the time-varying types of scene structures. Classification of saliency and gaze primitives and adaptive feature extraction schemes are introduced to effectively describe the correlations. We evaluate the proposed description with the task of attentive state estimation that classifies whether human observers concentrate on the displayed videos or not.

Chapter 5. Gaze Point Prediction based on Spatiotemporal Correlations

Based on the aforementioned two studies that separately focus on event-level spatiotemporal gaps and scene-level correlations, we finally introduce the description of overall spatiotemporal correlations. The framework with the PSDM is adopted here to deal with various categories of videos including unedited natural ones such as surveillance videos.

The main aim of this chapter is to describe how event-level spatiotemporal gaps can be influenced by scene-level correlations. To this end, we first introduce a model of gap structures that jointly deal with spatiotemporal gaps and local scene structures with help from the PSDM. Then, we statistically learn the modeled gap structures for each type of gaze primitives to involve the overall scene-level correlations between scene structures and gaze dynamics. In this chapter, we leverage our framework for gaze-point prediction from videos and evaluate if the proposed description can improve the performance of prediction.

Chapter 2

Modeling of Saliency Dynamics

2.1 Introduction

This chapter discusses how various visual events and scene structures in videos can be modeled in the proposed framework. Visual events involve semantic and physical aspects that have a different influence upon human gaze dynamics. Specifically, the semantics provide "why human observers watched visual events" since human observers often direct their eyes to regions with specific semantic categories. For example, human faces are known to attract our attention well, and several gaze behavior analyses take particular note of the faces in video scenes [Cerf et al., 2007, Subramanian et al., 2011]. On the other hand, the physical aspects are capable of explaining "how much human observers are attracted to visual events". As will be reviewed in the following section, some specific physical characteristics, i.e., saliency, form a conspicuous influence upon observers.

When taking into account of these aspects in the framework, one of the central issues arises since both aspects have diversity as posed in the previous chapter. We address this issue by particularly focusing on the physical dynamic aspects of visual events while sacrificing their semantics and modeling them simply and efficiently using primitive spatiotemporal patterns. Specifically, we propose a model named *saliency dynamics models* that can leverage the spatiotemporal patterns of salient regions in videos for the characterization of visual events. Since salient regions are capable of attracting attention, the proposed models allow us to directly handle the influences of visual events upon gaze dynamics.

In the following sections, we first present a brief history as to studies on the saliency. Then, we overview the modeling of saliency dynamics in Section 2.2 and present several examples of models in Section 2.3 and Section 2.4.

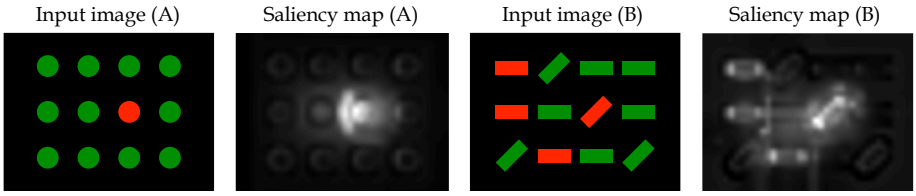

Figure 2.1: Input images and corresponding saliency maps with color, luminance and orientation channels generated by [Itti et al., 1998].

2.1.1 Visual Search and Saliency Maps

Assume a task to look for a unique stimulus from Input image (A) of Figure 2.1. With or without the instruction "the target is a red blob", we can effortlessly detect the red blob from the image. As for Input image (B), it may be more difficult to look for a unique stimulus than the previous example, but still we can detect the red skew bar easily. These tasks are called as the *visual search* to investigate visual attention mechanisms. The visual search requires subjects to judge if a specific target stimulus is present or not among other distractor stimuli and measures reaction times for the judgments, where the target and distractors are different in one or more basic image features (e.g., color and/or orientation in the above examples). The search can be basically classified into two types based on the relationships between targets and distractors: (1) feature search that adopts a target which can be distinguished from distractors by a unique feature like Input image (A), and (2) conjunction search with a target involving several features different from distractors like Input image (B).

Numerous psychophysical studies including the pioneering work undertaken by [Neisser and Beller, 1965] have aimed to clarify an attention mechanism via the visual search (see [Wolfe, 1998] as well as Review [Eckstein, 2011]). Particularly, one of the major theories is the feature integration theory (FIT) proposed by [Treisman and Gelade, 1980]. The FIT claims that (1) each single feature is processed by a unique module spatially in parallel and humans do not have to pay attention to a specific location during the feature search, and (2) the conjunction search requires to pay attention to a specific location to localize and integrate the results from multiple modules. Since the conjunction search requires a sequential procedure, reaction times to detect targets were believed to be proportional to the number of distractors (i.e., the set size effect [Palmer, 1994]) while the feature search requires constant reaction times regardless of the number of stimuli[(i)].

[(i)]This claim is basically no longer acceptable after [Wolfe, 1998] that has revealed that the dis-

Now, the focus of relevant research fields turns to how we can implement theories and models of visual attention in computer systems. The saliency map, first proposed by [Itti et al., 1998], is one of the representative implementations based on the FIT and a model of the shift in selective visual attention proposed by [Koch and Ullman, 1985]. The calculation of saliency maps begins with extracting several basic image features from an input image to construct a Gaussian pyramid for each image feature. Then, spatial contrasts of the image features between the center (fine) and surround (coarse) scales of the Gaussian pyramid, which are sometimes called as center-surround differences, are calculated and normalized at multiple scales. The obtained contrasts after normalization are referred to as a feature map. Finally, the feature maps are integrated over scales and image features to derive a saliency map (see [Itti et al., 1998] and Reviews [Borji and Itti, 2012, Kimura et al., 2013] for more detail). Figure 2.1 describes the corresponding saliency maps with color, luminance and orientation channels of Input images (A) and (B).

2.1.2 Various Extensions of Saliency Maps

Calculating center-surround differences

Although the original saliency map proposed by [Itti et al., 1998] is based on various theories of visual attention, recent studies to detect salient stimuli take a different approach to the extraction of center-surround differences. For example, [Gao and Vasconcelos, 2009] has introduced a measure of the center-surround differences based on the decision theory. Specifically, the degree of saliency in a certain location is high if distributions of feature values collected from small (center) and large (surround) windows around the location are easy to be discriminate. In addition, several models refer to the rarity of feature values at a certain location in a whole image, such as [Achanta et al., 2008, Cheng et al., 2011, Bruce and Tsotsos, 2009]. The models based on the rarity are sometimes called as *salient region detection*. Recently, a unified perspective of these center-surround differences has been presented in [Huang and Ahuja, 2012].

tributions of reaction times $\times$ the set size was unimodal.

Spatiotemporal extensions

There are many extensions to saliency maps besides the examples above. One of the promising directions is to introduce spatiotemporal saliency (also known as motion saliency) derived from spatiotemporal variations of images. A typical approach along this direction is to calculate a contrast of motion features such as optical flows [Tsotsos et al., 1995, Vijayakumar et al., 2001, Marat et al., 2009], flickers [Itti et al., 2003] as a channel of saliency. Several models that calculate rarity of feature values in a spatial patch can be extended to obtain spatiotemporal saliency by extracting features from spatiotemporal patches [Bruce and Tsotsos, 2009, Seo and Milanfar, 2009, Mahadevan and Vasconcelos, 2010]. Those patch-based approaches can involve longer-term motions than typical inter-frame motion features. In this context, several studies introduce temporal filters or parametric models to deal with the temporal decays of features [Zhang et al., 2009, Baldi and Itti, 2010]. Moreover, fusing spatial and spatiotemporal saliency is also an important issue and discussed in [Marat et al., 2009, Baldi and Itti, 2010].

The above approaches basically implicitly or explicitly suppress global motions such as camera motions from their calculation. Namely, global motions do not influence center-surround differences since these motions can be regarded as uniform in a certain local patch. On the other hand, the global motions themselves have the potential to attract attention as presented in [Yamada et al., 2010].

Semantic extensions

Another direction of extensions is to incorporate top-down (semantic) influences into the framework. Numerous extensions along this direction have been proposed so far: for example, tasks [Navalpakkam and Itti, 2005, Torralba et al., 2006, Peters and Itti, 2007, Borji et al., 2012], target characteristics [Frintrop et al., 2005, Navalpakkam and Itti, 2007] and object categories [Cerf et al., 2007, Judd et al., 2009, Borji, 2012]. In this context, learning-based approaches to associate saliency-related features with gaze points are often introduced [Peters and Itti, 2007, Judd et al., 2009, Borji, 2012, Borji et al., 2012][(ii)].

[(ii)] Note that a family of learning-based saliency maps and our gaze point prediction method proposed in Chapter 5 are different from a traditional saliency map that calculates center-surround differences of image features as an implementation of visual attention mechanisms. While the traditional saliency map deals with covert and overt attention (recall the performance of visual search tasks is evaluated by reaction times), the learning-based saliency is specialized to predict the overt attention since it is learned and evaluated with the points of gaze.

2.2 Modeling of Saliency Dynamics

2.2.1 Saliency of Visual Events

The modeling of saliency dynamics is aimed at leveraging spatiotemporal patterns of salient regions for the characterization of visual events. First, we present how various visual events appear in saliency maps using simple visual stimuli shown in Figure 2.2. In what follows, we refer to the local regions consisting of highly salient points in the maps as salient regions. We adopt saliency maps by [Itti et al., 1998] with the only luminance and motion channels since the input videos in Figure 2.2 consist of gray-scale images with motions. Moreover, the 3rd and 6th rows of the figure depict pixel-wise inter-frame differences of saliency maps to show how salient regions change due to visual events.

Object disappearances and appearances In Examples (1) and (2) in Figure 2.2, a white blob disappears and appears at the bottom-right of frames. As shown in the differences of saliency maps, the disappearances and appearances have an influence upon the degree of saliency not only at the regions of objects with the events but also those of other objects (i.e., the white blob at the top-left). It is because the saliency is generally normalized so that the amount of saliency in each frame is constant. Such interactions of saliency among objects are natural in practice since existing objects (e.g., the top-left blob) get or lose a chance to be looked at when a new event occurs (disappearances or appearances of the bottom-right blob).

Object translations and deformations Examples (3) and (4) describe translations and deformations (resizes) of displayed objects. As shown in the saliency maps, these events bring the translations and deformations of salient regions, respectively. Namely, if the events are caused by distinct objects like these examples, the changes in the objects are reflected to those of salient regions in a homogeneous manner. In addition, we can also find interactions of saliency in Example (4); the white blob in the top-left obtains saliency since that in the bottom-right gets small.

Scene changes The scene change from Example (4) to Example (5) can be mostly explained by the disappearances and appearances of objects. Since the saliency generally takes into account of local center-surround differences,

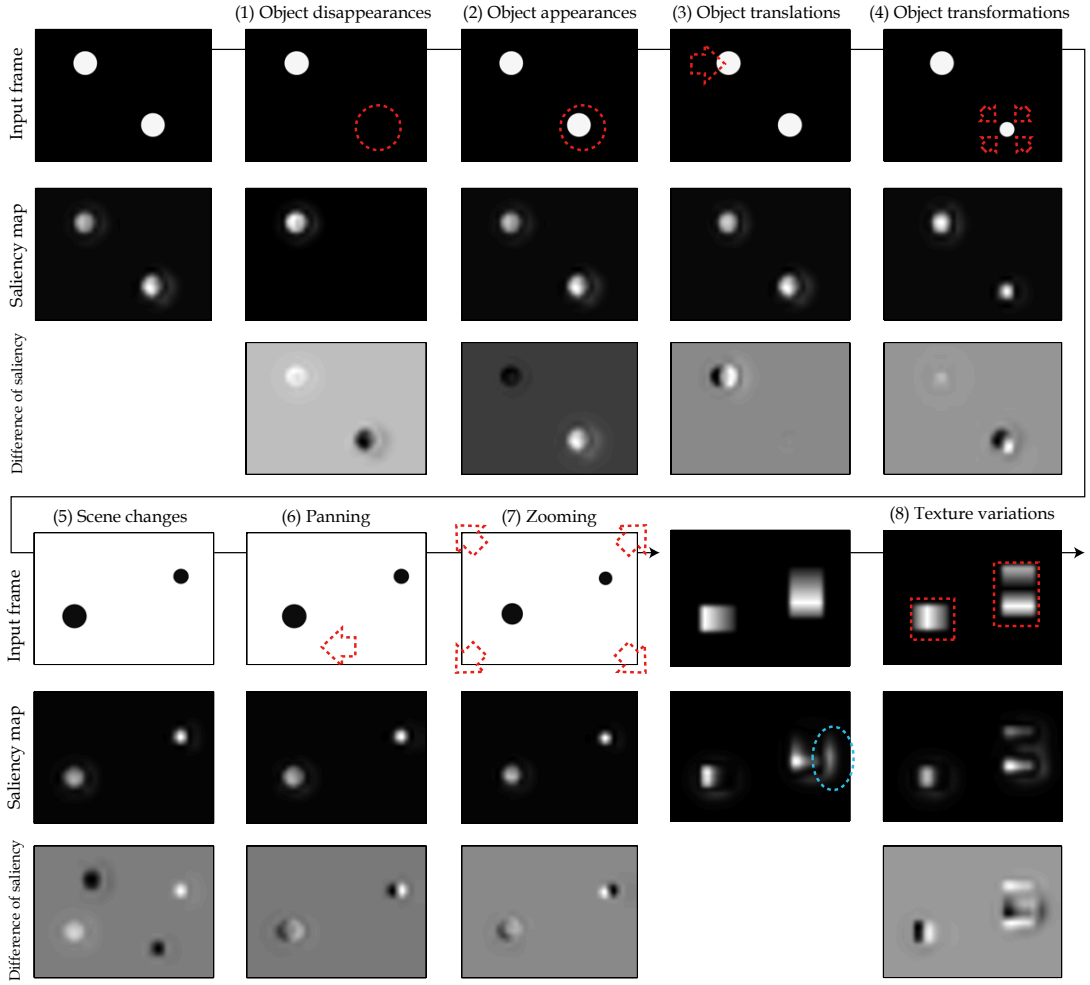

Figure 2.2: Various visual events and corresponding saliency maps with luminance and motion channels generated by [Itti et al., 1998]. The 3rd and 6th rows describe the pixel-wise differences between successive saliency maps.

the global change in luminance from black to white does not influence saliency maps even if the maps have a motion channel.

Panning and zooming Camera motions such as panning and zooming in Examples (6) and (7) can be regarded as the same as the variations of objects such as translations and deformations. When existing objects disappear or new objects appear due to the camera motions, they can be also described by the disappearances and appearances of objects explained above.

Texture variations Finally, Example (8) shows texture variations (variations of luminance gradations) in certain local spatiotemporal patches. The variation of the left texture appears in the saliency maps as a translation of regions. On the other hand, the right texture appears as the combination of

resize and appearance of two salient regions. In other words, the variations of two salient regions characterize the right texture variation events on the whole. Note that there is another salient region outside the patches as shown by the blue oval line, which is generated due to the contrast of luminance along the texture patch. In this manner, salient regions sometimes exist in backgrounds as a result of foreground events.

2.2.2 Modeling Saliency Dynamics with Saliency Primitives

Now we introduce an overview of the modeling of saliency dynamics with saliency primitives. In what follows, the definition and basic concept of saliency dynamics models are summarized first. Then, we pose several options for the modeling of saliency primitives to effectively deal with visual events. Finally, we introduce two typical categories of videos that require different options as a modeling target.

The definition and basic concept of saliency dynamics models

Assume that a sequence of saliency maps is obtained from video frames. Since various visual events can occur in the videos and they are reflected in the dynamic changes of salient regions as shown in the previous section, the obtained sequence can contain salient regions dynamically changing over time. In addition to the individual changes of salient regions, the dynamics that the changing patterns follow as well as the number of the salient regions also have a temporal variation as overall scene structures change over time. In this study, we refer to such dynamic changes provided by salient regions as *saliency dynamics*.

The basic concept of our saliency dynamics models is to introduce primitive spatiotemporal patterns of salient regions as a unit of the modeling. The primitive patterns, which we refer to as *saliency primitives*, describe the changes of salient regions caused by various visual events. Furthermore, the sets of primitives can characterize overall scene structures consisting of multiple visual events simultaneously occurring in a certain temporal interval. By modeling saliency primitives appropriately and learning them from a set of videos, we can describe how visual events and scene structures influence human gaze dynamics based on the saliency dynamics models efficiently configured for the given videos.

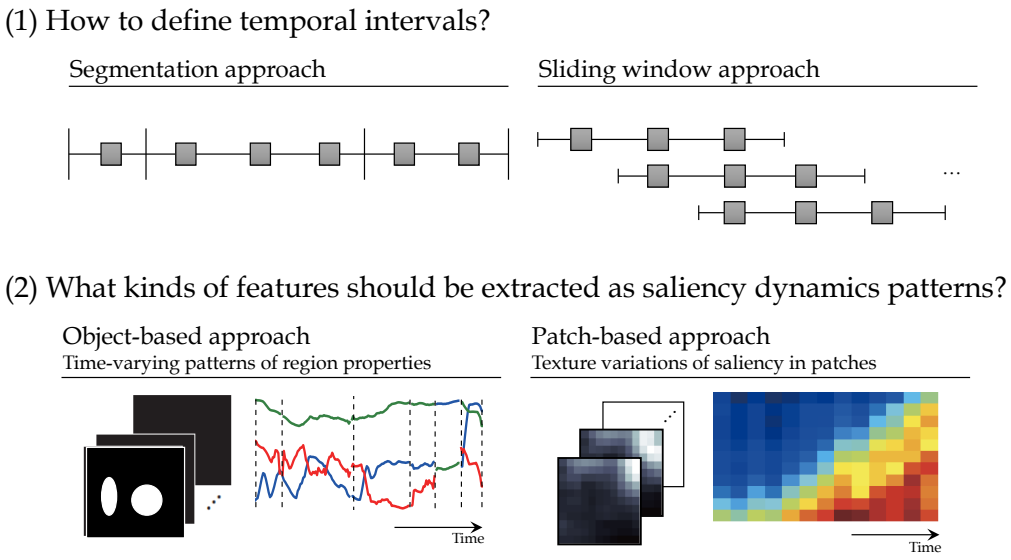

Figure 2.3: Options for the modeling of saliency primitives.

Options for the modeling of saliency primitives

There are several options for the modeling of saliency primitives to effectively deal with visual events. The first option is about how to define a temporal interval to extract spatiotemporal patterns for saliency primitives (Figure 2.3 (1)). For this, we can introduce two approaches to the definitions called *segmentation* and *sliding windows*. They are well known in the field of time series pattern recognition and data mining (see Review [Ratanamahatana et al., 2005]). The segmentation approach looks for a set of points (segmentation points) where the temporal intervals split. This approach can explicitly deal with scene change events while it has difficulty in detecting segmentation points so as not to split spatiotemporal patterns incorrectly. On the other hand, the sliding-window approach slides a fixed-length window from the beginning to the end of sequences with an overlap, and conducts a certain procedure in the temporal intervals defined by each window. This approach can avoid splitting spatiotemporal patterns incorrectly thanks to the redundant representation by the overlap although it is known that

the clustering of pattens fails if two patterns in successive sliding windows are extremely similar [Keogh et al., 2003]. Also, we cannot deal with scene change events explicitly when adopting the sliding-window approach.

Given a temporal interval, the second option is about what kinds of features should be extracted as saliency dynamics patterns in the interval (Figure 2.3 (2)). If we want to take particular note of variations in a sole salient region, for example when we deal with events caused by distinct objects (Examples (1) to (7) in Figure 2.2), the properties of regions such as positions, shapes and the degree of saliency can be explicitly utilized. On the other hand, when we deal with a more general variation including texture variations, the changes of a sole salient region cannot always describe the whole variations as shown by Example (8) in Figure 2.2. In this case, it is effective to describe the patterns of one or more salient regions jointly and implicitly as parts of the texture variations of saliency in a certain spatiotemporal patch. We refer to these two approaches as *object-based* and *patch-based* approaches in what follows.

The third option is how to represent the extracted patterns in the saliency primitives (Figure 2.3 (3)). As reviewed in Section 1.2.1, the representations are dropped into two categories: direct and model-based approaches. While model-based representations like SLDS [Bregler, 1997, North et al., 2000, Li et al., 2002] and dynamic textures [Doretto et al., 2003, Chan and Vasconcelos, 2008, Ravichandran et al., 2009] are aimed at representing patterns efficiently with small number of parameters, we need to define a suitable model for the given patterns. On the other hand, direct representations are model-free and they can deal with any kinds of patterns although they take an ingenuity to avoid diversity and noise in the patterns.

In conclusion, the first and second options are related with what types of visual events should be considered. Temporal intervals should be defined based on whether we give an importance on scene changes or not. The second option should be chosen based on if we consider visual events caused by the only distinct objects or more general ones. On the other hand, the third option is about how to obtain efficiency when describing primitives. In every case, the modeling of primitives is aimed at describing a greater variety of dynamic changes compared to a simple motion analyses such as optical flows. While the optical flows generally provide the power and orientation of motions in each pixel, our saliency primitives describe not only translations of regions but their deformations and texture variations of saliency that appear in a certain temporal interval.

Table 2.1: Differences between OSDMs and PSDMs.

	OSDM (Section 2.3)	PSDM (Section 2.4)
Applicable video types	Intentionally-designed videos	Unedited natural videos
Definition of temporal intervals	Segmentation	Sliding-windows
Features to be extracted	Properties of regions (object-based)	Textures (patch-based)
Representations of primitives	Model-based	Direct

Video categories and proposed saliency dynamics models

Finally, we introduce specific saliency dynamics models based on the options presented above. For guidance to choose the options, we here introduce two categories of videos that individually tend to contain specific types of visual events.

Intentionally-desgined videos. The videos taken with a certain objective, e.g., TV commercial films and movies, are designed to attract observers' attention on intended objects (logos, products and so on), and thus the limited number of objects can be shown simultaneously in a certain temporal interval. These objects are mostly highly salient since they are designed to make their appearance distinct relative to their surrounds. In addition, they often involve frequent scene changes to give much information to observers. We can witness more designed videos in our daily life: for example, navigation interfaces placed in airports and shopping malls. These videos often adopt a few limited layouts of intended objects for the sake of usability.

Unedited natural videos. Videos recorded under uncontrolled situations without intentions do not always contain the limited number of objects with high saliency. For example, plain natural sceneries sometimes contain less objects. On contrary, surveillance videos with human crowds contain massive objects. Note that visual events are often regarded as texture variations when analyzing such natural videos: e.g., [Chan and Vasconcelos, 2008, Zhan et al., 2008, Ravichandran et al., 2009]. Moreover, unedited videos basically have less scene changes.

These two categories of videos require different options when modeling the saliency primitives. We thus propose two models of saliency dynamics which are individually suitable for those categories. Specifically, we refer to the model

for intentionally-designed videos as *object-based saliency dynamics models (OSDM)* and for unedited natural videos as *patch-based saliency dynamics models (PSDM)*. The OSDM is aimed at describing visual events caused by distinct objects as well as scene changes. On the other hand, the PSDM introduces the modeling of saliency primitives suitable for a greater variety of local events including texture variations. The differences of the options adopted in these two models are summarized in Table 2.1. Note that they are not the unique models against the two video categories. For example, we can introduce model-based representation of primitives in the PSDM, like a family of dynamic textures [Doretto et al., 2003, Chan and Vasconcelos, 2008, Ravichandran et al., 2009].

2.2.3 Notations

Let $\boldsymbol{p} = (x, y) \in \Omega$ be a 2-d point in a frame of videos, where $\Omega \subset \mathbb{R}_+^2$ is a spatial domain corresponding to the frame. We particularly use $\boldsymbol{p}_t = (x_t, y_t)$ if we specify a certain point at frame $t \in \mathbb{N}$. The saliency maps are denoted as $S : \Omega \to \mathbb{R}_+$, where the degree of saliency at point $\boldsymbol{p}$ is $S(\boldsymbol{p})$. Above all, we specify the saliency map at frame t as S_t and the local regions $\Omega' \subseteq \Omega$ of S_t as $S_{(\Omega',t)}$ (i.e., $S_t = S_{(\Omega,t)}$). Then, a sequence of saliency maps obtained from a video can be denoted as an ordered set, $\mathcal{S} = (S_1, \ldots, S_T)$, where T is the number of frames. If we introduce a local spatiotemporal patch defined as $\Omega' \times \mathcal{T}$ where $\mathcal{T} \subseteq [1, T]$, the local spatiotemporal volume in the patch is denoted as $\mathcal{S}_{\Omega' \times \mathcal{T}} = \left(S_{(\Omega',\min(\mathcal{T}))}, \ldots, S_{(\Omega',\max(\mathcal{T}))}\right)$ where $\min(\mathcal{T})$ and $\max(\mathcal{T})$ are lower and upper bounds of $\mathcal{T}$, respectively.

2.3 Object-Based Saliency Dynamics Model

2.3.1 Overview of the Model

The OSDM is aimed at modeling saliency dynamics provided by intentionally-designed videos such as TV commercial films. These videos contain visual events from distinct objects (e.g., translations of objects) and frequent scene changes. As summarized in Table 2.1, the options chosen in the OSDM are as follows:

1. We segment a whole video into small temporal intervals to deal with scene change events explicitly.

2. We parametrically describe the properties of salient regions such as positions, contour shapes and the degree of saliency so that we can handle visual events caused by distinct objects.

3. We parametrically model the spatiotemporal patterns of those properties by saliency primitives.

For simplicity, we assume that the scene structures in each temporal interval, which are characterized by the spatiotemporal patterns and the number of salient regions, are independent. Then, a key problem is how to detect segmentation points that give reasonable intervals to model the patterns in each interval by saliency primitives accurately. In this section, we first introduce a formulation of the OSDM, and then we propose an approach to a model estimation including a segmentation technique.

2.3.2 Formulation

We first assume that videos are segmented into a sequence of K temporal intervals, $\mathcal{I} = (I_1, \ldots, I_K)$. Saliency maps in interval $I_k = [i_{k1}, i_{k2}]$, $\{S_t \mid t \in I_k\}$ individually contain C_k salient regions, where the spatiotemporal pattern of the c-th region is described by a sequence of multidimensional vectors, $\Theta_k^{(c)} = \left(\boldsymbol{\theta}_{i_{k1}}^{(c)}, \ldots, \boldsymbol{\theta}_{i_{k2}}^{(c)}\right)$ ($\boldsymbol{\theta}_t^{(c)} \in \mathbb{R}^J$ denotes the properties of the c-th region in frame t). Then, the saliency dynamics in interval I_k, which characterize a scene structure in the interval, can be represented by a set of patterns, $\boldsymbol{\Theta}_k = \{\Theta_k^{(1)} \ldots, \Theta_k^{(C_k)}\}$.

We describe each pattern $\Theta_k^{(c)}$ by a single saliency primitive modeled in a parametric manner. Since distinct objects in our videos of interests mostly behave naturally to attract our attention, the corresponding patterns of salient regions seem to follow some dynamical systems. We thus define saliency primitive $D_k^{(c)}$ identified to $\Theta_k^{(c)}$ by a first-order multivariate autoregressive model (AR model) as a family of LDSs:

$$\boldsymbol{\theta}_t^{(c)} = M_k^{(c)} \boldsymbol{\theta}_{t-1}^{(c)} + \boldsymbol{b}_k^{(c)} + \boldsymbol{v}_t, \tag{2.1}$$

where $M_k^{(c)}$ is a $J \times J$ transition matrix, $\boldsymbol{b}_k^{(c)}$ is a J-dimensional bias vector, $\boldsymbol{v}_t$ is a J-dimensional noise vector modeled by a Gaussian distribution $\mathcal{N}(0, Q_k^{(c)})$. Namely, saliency primitive $D_k^{(c)}$ has $M_k^{(c)}, \boldsymbol{b}_k^{(c)}, Q_k^{(c)}$ as parameters.

In terms of describing complex patterns by the switches of simple models, the proposed OSDM is similar to the switching linear dynamical systems (SLDS) adopted in [Bregler, 1997, North et al., 2000, Li et al., 2002]. Comparing to the

SLDS, the PSDM introduces a set of AR models to describe dynamics in a certain interval, and thus it has an advantage in describing the situations where the number of elements (objects) providing the dynamics change over time.

2.3.3 Extraction and Modeling of Salient Regions

When introducing OSDMs, we first need to model $\boldsymbol{\theta}_t^{(c)}$ so as to describe properties of salient regions (where c is an ID given to a salient region and t is a frame ID). Among the properties, shape contours (region boundaries) can be describe by Snakes [Kass et al., 1988] and Level Sets [Cremers, 2006], for example. In addition, Active Appearance Models (AAM) [Cootes et al., 2001] can deal with shapes and appearances (textures) inside the shapes jointly by deriving appearances for warped shapes, where both are efficiently represented via the principal component analysis. However, the performance of AAMs greatly depends on feature-point selection manually conducted in the training of models.

In this study, we model salient regions in a frame by the Gaussian mixture model (GMM). That is, each salient region is modeled by a single Gaussian component. This modeling sacrifices representation of a detailed contour and texture of the regions. Instead, the GMM allows us to describe locations, approximate shapes, and the degree of saliency of the regions with mean vectors, covariance matrices and a weight factor of the GMMs, respectively. In addition, the GMM can represent the interactions of saliency among objects as mentioned in Section 2.2.1 thanks to the weight factors. Let us denote the mean vector, the covariance matrix and the weight of the c-th Gaussian component as $\boldsymbol{\mu}_t^{(c)}, \Sigma_t^{(c)}, \phi_t^{(c)}$, where the number of components is C_k.

To estimate parameters of the GMM, we first generate massive samples along probability distribution $\bar{S}_t = S_t(\boldsymbol{p}) / \sum_{\boldsymbol{p}\in\Omega} S_t(\boldsymbol{p})$. Then, we estimate the parameters from the samples via the standard expectation-maximization (EM) algorithm [Bishop, 2006]. The EM algorithm performs as a local optimization and has a strong dependency for initial values. In addition, estimated parameters need to have a continuous change over time when we model them by linear AR models. Thus, we give estimated parameters at a certain frame as initial inputs in the next frame, where the initials of means at the beginning frame of intervals are locations of samples chosen randomly, those for covariance matrices are set to be diagonal and calculated from all the samples regardless of component IDs, and those for weight factors are uniform.

As we represent series of salient regions by GMMs in a frame-wise manner, we obtain spatiotemporal patterns of C_k regions in interval $I_k = [i_{k1}, i_{k2}]$ by tracking Gaussian components at the spatiotemporal nearest neighbor in the previous frame from the beginning of the interval to the end. Specifically, we track the c-th component in i_{k1} to obtain the patterns of mean vectors, covariance matrices and weight factors, $(\boldsymbol{\mu}_{i_{k1}}^{(c)}, \ldots, \boldsymbol{\mu}_{i_{k2}}^{(c)})$, $(\boldsymbol{\sigma}_{i_{k1}}^{(c)}, \ldots, \boldsymbol{\sigma}_{i_{k2}}^{(c)})$, $(\phi_{i_{k1}}^{(c)}, \ldots, \phi_{i_{k2}}^{(c)})$, respectively, where $\boldsymbol{\sigma}_t^{(c)} \in \mathbb{R}^3$ consists of two variances and a covariance of $\Sigma_t^{(c)}$. Finally, we denote the properties of the c-th region at frame t as $\boldsymbol{\theta}_t^{(c)} = ((\boldsymbol{\mu}_t^{(c)})^{\mathrm{T}}, (\boldsymbol{\sigma}_t^{(c)})^{\mathrm{T}}, \phi_t^{(c)})^{\mathrm{T}} \in \mathbb{R}^6$, which constitutes spatiotemporal pattern $\Theta_k^{(c)}$.

2.3.4 Identification of Saliency Primitives and Segmentation

Problem settings

The OSDM introduces temporal interval sequence $\mathcal{I} = (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_K)$ to deal with time-varying scene structures that characterize scene change events. Each interval contains a set of spatiotemporal patterns of salient regions, where they are supposed to be identified by saliency primitives defined in Equation (2.1).

This section introduces a model estimation consisting of the segmentation of input saliency maps to derive appropriate interval sequence $\mathcal{I}$ and the description of spatiotemporal pattern $\Theta_k^{(c)}$ by saliency primitive $D_k^{(c)}$. Segmentation $\mathcal{I}$ should be given so as to identify $D_k^{(c)}$ with small identification costs (i.e., fitting errors) to $\Theta_k^{(c)}$ in interval I_k. On the other hand, $\mathcal{I}$ should be given preliminarily when identifying primitives to spatiotemporal patterns and evaluate costs.

To address this problem, we first generate many temporal interval candidates and select an appropriate segmentation based on identification costs of saliency primitives in each candidate. Specifically, we first generate hierarchical structures of interval candidates based on a scale-space representation of inter-frame differences of saliency maps (Figure 2.4 (2-1)) and fit GMMs to extract salient regions and identify saliency primitives to their patterns in each interval candidate (Figure 2.4 (1-1) and (1-2)). Then, we evaluate segmentation points defined by two successive interval candidates based on identification costs of the primitives and derive a whole segmentation (Figure 2.4 (2-2)). As a consequence, we can conduct the segmentation based on the identification costs of saliency primitives.

In what follows, we first propose a method to calculate identification costs from a set of saliency primitives in a given interval. Then, we present a method for segmentation based on the calculated identification costs.

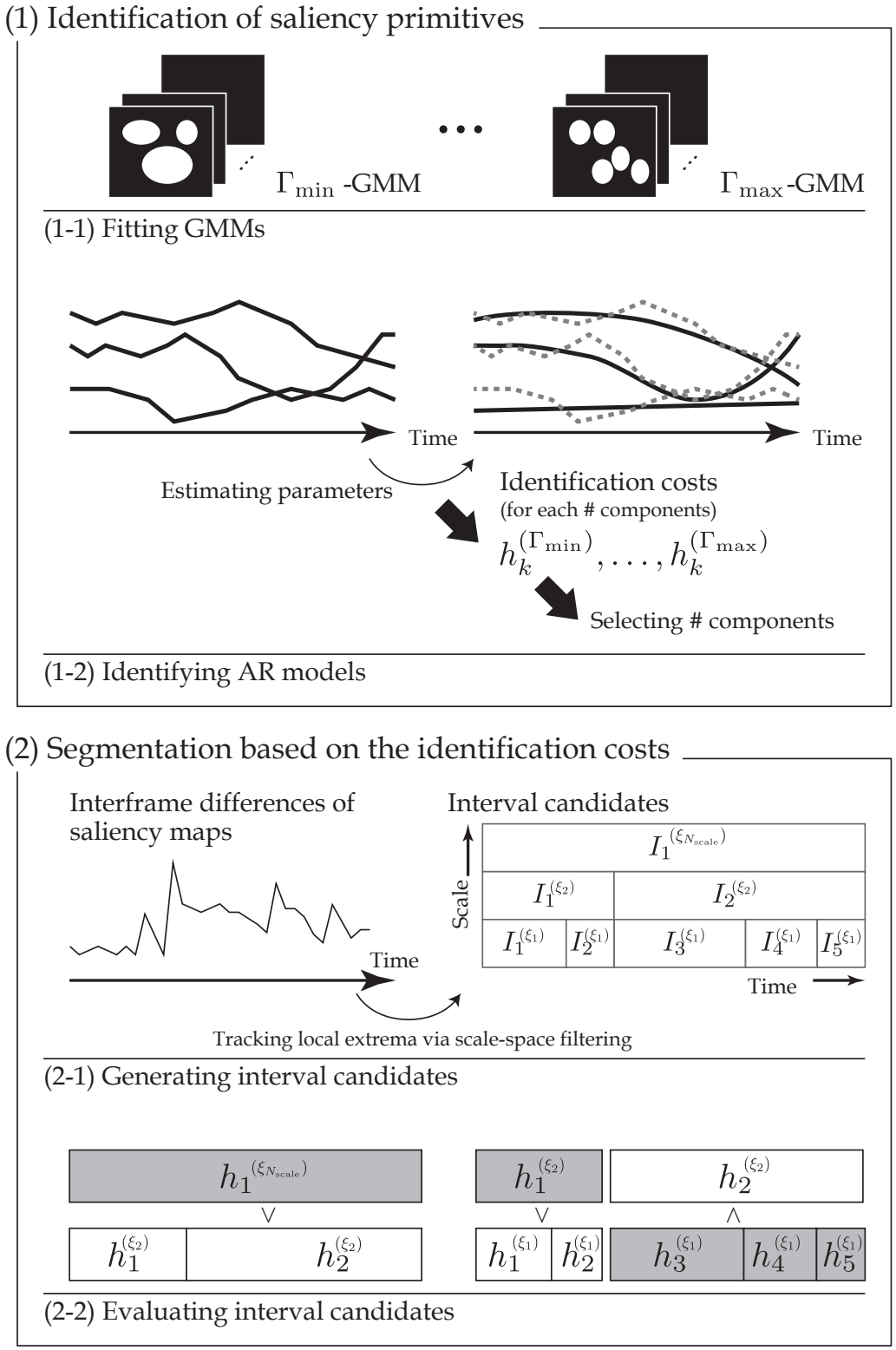

Figure 2.4: Estimation algorithm for the object-based saliency dynamics models.

Identification of saliency primitives (Figure 2.4 (1-1) and (1-2))

Given a certain interval, $I_k = [i_{k1}, i_{k2}]$, the identification procedure consists of estimating the number of components in GMM and at the same time identifying saliency primitives (i.e., AR models) to each spatiotemporal pattern in the sequence of saliency maps $(S_{i_{k1}}, \ldots, S_{i_{k2}})$ (Figure 2.4 (1)). We first set a range to the number of components, $\{\Gamma_{\min}, \Gamma_{\max}\}$ and fit $\Gamma_{\min}, \ldots, \Gamma_{\max}$-component GMMs individually to S_t via the procedure in Section 2.3.3. We describe the spatiotemporal pattern of the c-th of $\Gamma \in \{\Gamma_{\min}, \ldots, \Gamma_{\max}\}$ regions in the k-th interval as $\Theta_k^{(c,\Gamma)}$.

Let us denote the saliency primitive identified to $\Theta_k^{(c,\Gamma)}$ as $D_k^{(c,\Gamma)}$. As defined in Equation (2.1), $D_k^{(c,\Gamma)}$ has a set of parameters consisting of transition matrix $M_k^{(c)}$, bias vector $\boldsymbol{b}_k^{(c)}$ and error covariance matrix $Q_k^{(c)}$ (in what follows, we omit subscript Γ without loss of generality). $M_k^{(c)}$ and $\boldsymbol{b}_k^{(c)}$ can be basically estimated by minimizing a prediction error from $\boldsymbol{\theta}_{t-1}^{(c)}$ to $\boldsymbol{\theta}_t^{(c)}$. However, positions, shapes and the degree of saliency of salient regions, which are described by elements of $\Theta_k^{(c)}$, sometimes perform high correlation to each other. In that case, we cannot estimate parameters correctly due to a multicollinearity problem.

Thus, we adopt the ridge regression problem that adds a regularization term $||\Lambda_k^{(c)}||_{\mathrm{F}}^2$ on $J \times (J+1)$ matrix $\Lambda_k^{(c)} = [M_k^{(c)} \mid \boldsymbol{b}_k^{(c)}]$ ($||\cdot||_{\mathrm{F}}$ is a Frobenius norm of matrices). Let us denote $X_k^{(c)} = [(\boldsymbol{\theta}_{i_{k1}}^{(c)}, \ldots, \boldsymbol{\theta}_{i_{k2}-1}^{(c)})^{\mathrm{T}} \mid \mathbf{1}_{T\times 1}]^{\mathrm{T}}$ ($(J+1) \times T$ matrix), $Y_k^{(c)} = [\boldsymbol{\theta}_{i_{k1}+1}^{(c)}, \ldots, \boldsymbol{\theta}_{i_{k2}}^{(c)}]$ ($J \times T$ matrix), where $T = i_{k2} - i_{k1}$, $\mathbf{1}_{T\times 1}$ is a T-dimensional all-ones vector. Then, the problem is formalized as follows:

$$\hat{\Lambda} = \arg\min_{\Lambda_k^{(c)}} \left\{ \|Y_k^{(c)} - \Lambda_k^{(c)} X_k^{(c)}\|_{\mathrm{F}}^2 + \lambda \|\Lambda_k^{(c)}\|_{\mathrm{F}}^2 \right\}. \tag{2.2}$$

We can derive $\hat{\Lambda}$ by analytically solving Eq. (2.2):

$$\hat{\Lambda} = \left(\lambda \boldsymbol{I} + (X_k^{(c)})^{\mathrm{T}} X_k^{(c)}\right)^{-1} (X_k^{(c)})^{\mathrm{T}} Y_k^{(c)}.$$

Note that we set regularization parameter λ such that $||\Lambda_k^{(c)}||_{\mathrm{F}}$ is smaller than a preliminarily defined threshold.

Once we estimate parameter $\hat{\Lambda}$, we can generate spatiotemporal pattern $\hat{\Theta}_k^{(c)} = (\hat{\theta}_{i_{k1}}^{(c)}, \ldots, \hat{\theta}_{i_{k2}}^{(c)})$ from given initial value $\boldsymbol{\theta}_{i_{k1}}^{(c)}$ of original pattern $\Theta_k^{(c)} = (\theta_{i_{k1}}^{(c)}, \ldots, \theta_{i_{k2}}^{(c)})$. We then calculate error covariance matrix $Q_k^{(c)}$ by modeling the distribution of errors between original and generated patterns, $\theta_t^{(c)} - \hat{\theta}_t^{(c)}$ by a normal distribution: $\theta_t^{(c)} - \hat{\theta}_t^{(c)} \sim \mathcal{N}(0, Q_k^{(c)})$. In addition, we can calcu-

late a negative log likelihood (NLL) score $h_k^{(c)}$ by evaluating the errors, $h_k^{(c)} = -\sum_{t=i_{k1}}^{i_{k2}} \log P(\theta_t^{(c)} - \hat{\theta}_t^{(c)}; 0, Q_k^{(c)})$.

As a result of the procedure above, we have a set of saliency primitives $\{D_k^{(c,\Gamma)} \mid c = 1\ldots,\Gamma\}$ and corresponding NLL scores $\{h_k^{(c,\Gamma)} \mid c = 1\ldots,\Gamma\}$ for each $\Gamma \in \{\Gamma_{\min},\ldots,\Gamma_{\max}\}$. To determine the number of components that is the most suitable for introducing saliency primitives at the k-th interval, we first evaluate the worst fit of primitives for each Γ, $h_k^{(\Gamma)} = \max\{h_k^{(c,\Gamma)} \mid c = 1\ldots,\Gamma\}$. As Γ increases from $\Gamma_{\min}$ to $\Gamma_{\max}$, NLL score $h_k^{(\Gamma)}$ decreases until the fitness of saliency primitives becomes sufficiently good. We thus define $\hat{\Gamma}_k$ as the point where the NLL scores stop decreasing. Finally, we obtain primitive set $\{D_k^{(1)}\ldots D_k^{(\hat{\Gamma}_k)}\}$ from spatiotemporal patterns in temporal interval I_k, $\boldsymbol{\Theta}_k = \{\Theta_k^{(1)},\ldots,\Theta_k^{(\hat{\Gamma}_k)}\}$, where the identification cost of primitives is given as $h_k = h_k^{(\hat{\Gamma}_k)}$.

Segmentation based on the scale-space analysis (Figure 2.4 (2-1) and (2-2))

Video segmentation is a well-known problem to detect scene change events in visual content analyses as reviewed in [Cotsaces et al., 2006]. Our segmentation technique presented below is aimed at detecting the scene changes with the object to describe saliency dynamics patterns in each interval accurately by a set of saliency primitives.

The basic idea is that we generate multiple interval candidates base on the scale-space analysis [Witkin, 1983] and evaluate the segmentation points between successive interval candidates based on the identification costs of primitives (Figure 2.4 (2-1)). Specifically, we first fit $\Gamma_{\max}$-component GMM for each of the frames and concatenate mean vectors $\boldsymbol{\mu}_t^{(\mathrm{cat})} = ((\boldsymbol{\mu}_t^{(1)})^{\mathrm{T}},\ldots,(\boldsymbol{\mu}_t^{(\Gamma_{\max})})^{\mathrm{T}})^{\mathrm{T}}$ as a $2\cdot\Gamma_{\max}$-dimensional feature vector for S_t. We then calculate inter-frame difference $f_t \in \mathbb{R}$ between successive saliency maps S_{t-1}, S_t as $f_t = |\boldsymbol{\mu}_t^{(\mathrm{cat})} - \boldsymbol{\mu}_{t-1}^{(\mathrm{cat})}|$. Thanks to the description of frames based on the GMM, f_t can be sensitive to appearance and disappearance events of objects. Then, we use a sequence of inter-frame differences, $\boldsymbol{f} = (f_1,\ldots,f_T)$, as an input. We convolve a series of Gaussian functions with smoothing scales $\{\xi_1,\ldots,\xi_{N_{\mathrm{scale}}}\}$ $(\xi_{n-1} < \xi_n)$, let's say $Gauss^{(\xi_n)}$, to sequence $\boldsymbol{f}$ and obtain a scale-space representation $\boldsymbol{f}^{(\xi_n)} = \boldsymbol{f} * Gauss^{(\xi_n)}$, where $*$ denotes a convolution operation. By tracking local extreme points (i.e., inflection points of saliency map sequences) in a set of outputs $\{\boldsymbol{f}^{(\xi_1)},\ldots,\boldsymbol{f}^{(\xi_{N_{\mathrm{scale}}})}\}$ with changing the smoothing scales from $\xi_{N_{\mathrm{scale}}}$ to ξ_1, we can obtain a hierarchical structure of the points since new points can appear as decreasing smoothing scales due

to the causality of the Gaussian function. For simplicity of discussions, we set $\{\xi_1, \ldots, \xi_{N_{\text{scale}}}\}$ so as to obtain new local extreme points for every scale variation $\xi_n \rightarrow \xi_{n-1}$. In addition, we set the maximum scale $\xi_{N_{\text{scale}}}$ so as not to contain any local extreme point.

Given a local extreme point at certain scale ξ_n, we look for the corresponding point at scale ξ_1 by tracking the point from ξ_n to ξ_1 and use the point as one of the segmentation points at ξ_n. Then, we deal with segments defined by successive segmentation points as interval candidates. We denote the interval candidates generated at scale ξ_n as $\hat{\mathcal{I}}^{(\xi_n)} = (\hat{I}_1^{(\xi_n)}, \ldots, \hat{I}_{K_{\xi_n}}^{(\xi_n)})$. For each interval, a set of saliency primitives are identified individually with spatiotemporal patterns in $\hat{I}_k^{(\xi_n)}$ and identification cost $h_k^{(\xi_n)}$ is given to the interval based on the procedure presented in the previous section.

After obtaining identification costs for all the interval candidates, we can evaluate segmentation points (Figure 2.4 (2-2)). Let us introduce a subsequence of $\hat{\mathcal{I}}^{(\xi_{n-1})}$ at scale ξ_{n-1}, $\hat{\mathcal{I}}^{(\xi_{n-1})}\,|_{(j,j+l)} = (I_j^{(\xi_{n-1})}, \ldots, I_{j+l}^{(\xi_{n-1})})$, which defined in the same interval as candidate interval $\hat{I}_k^{(\xi_n)}$ at scale ξ_n. In the segmentation, we choose one of $\hat{I}_k^{(\xi_n)}$ and $\hat{\mathcal{I}}^{(\xi_{n-1})}\,|_{(j,j+l)}$ based on the identification costs (Figure 2.4 (2-2)). Specifically, we split the interval if $h_k^{(\xi_n)} \geq \sum_{j'=j}^{j+l} h_{j'}^{(\xi_{n-1})}$. By recursively conducting the judgements from $\xi_{N_{\text{scale}}}$ to ξ_1, we can obtain an appropriate segmentation to describe spatiotemporal patterns with saliency primitives.

2.3.5 Examples

This section introduces examples of model estimation results with actual intentionally-designed videos. Specifically, we adopted 12 TV commercial films of 15 sec length stored at 30 fps. These videos are designed to contain several distinct objects generating various visual events and scene changes over time. Before fitting the model, we first resized videos into 80×60 pixel resolution for the sake of computation speed. As for an input saliency map, we adopted the graph-based visual saliency [Harel et al., 2007], where the features include luminance, color, edge orientations and motions (contrasts in the amplitudes of pixel-level shifts). When estimating the number of GMM components, we approximated saliency maps with 20000 samples, where the samples were collected from the positions at which the degree of saliency was higher than 90 percentile in the distributions of saliency for each frame. The procedure above was aimed at avoiding capturing regions with the lower saliency other than distinct objects when fitting the

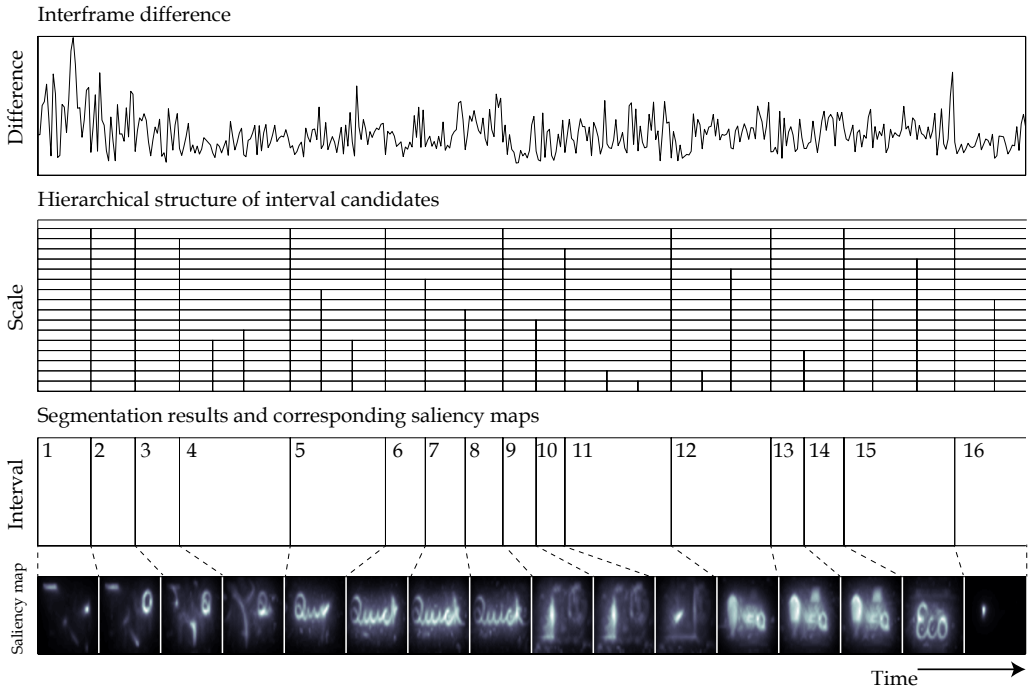

Figure 2.5: Example of segmentation results. 1st row: a sequence of inter-frame differences. 2nd row: hierarchical structures of interval candidates generated by the scale-space analysis of the inter-frame differences in the 1st row. 3rd row: segmentation results consisting of the selected intervals from the candidates in the 2nd row. the images below depict saliency maps at the beginning frame of each interval. The images used in this figure was provided by courtesy of Panasonic Corporation.

GMMs. In addition, we assumed there were only several objects in each frame of the videos and empirically set $\Gamma_{\min} = 1, \Gamma_{\max} = 8$. Under these settings, the number of intervals, K, was estimated at $11 \leq K \leq 19$ for any video (mean: 15.7, SD: 2.2). The number of primitives (i.e., salient regions) in each interval, C_k, was estimated at $2 \leq C_k \leq 5$ for any scene (mean: 2.8, SD: 0.7).

A selected example of segmentation results is depicted in Figure 2.5. Although many peaks were found in the inter-frame differences of saliency maps in the 1st row, the final segmentation in the 3rd row contained several scene change events such as the 4th to 5th, 8th to 9th, 11th to 12th, 14th to 15th and 15th to 16th intervals. In addition, appearance events of new objects also contribute to the switches of scene structures such as the 1st to 2nd and the 12th to 13th intervals.

Extracted spatiotemporal patterns of salient regions corresponding to Figure 2.5 are shown in the left of Figure 2.6. Obviously, the extracted patterns contain large noises. One of the reasons is the definition of saliency; saliency maps

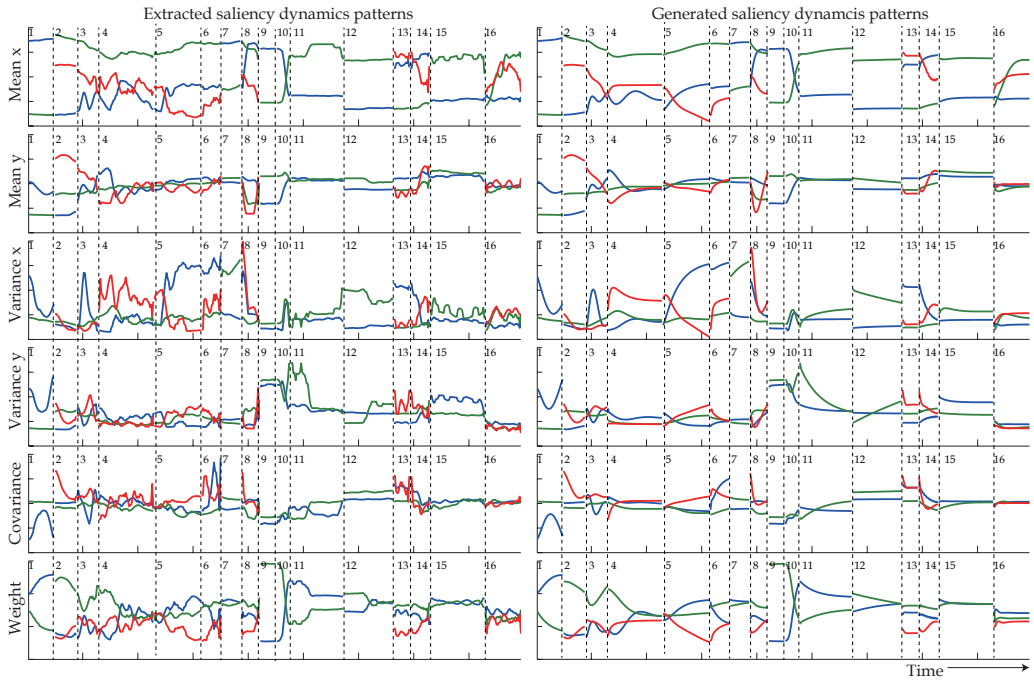

Figure 2.6: Extracted spatiotemporal patterns of salient regions (left of the figure) and generated patterns from the identified saliency primitives (right). Lines of the same colors in each interval between left and right of the figure indicate the same saliency primitive. Each row shows different properties of regions.

are generally obtained frame by frame and represent the degree of saliency at each point in a frame, and thus the point in the same object can obtain different saliency if the surrounding objects generate visual events as discussed in Section 2.2.1. Another reason is the instability in the fitting of GMMs. When salient regions are too large to model by a single Gaussian component, the proposed model introduce several components to represent the regions such as the 5th, 6th and 7th intervals, which sometimes makes the fitting unstable.

The right of Figure 2.6 depicts the generated patterns from identified saliency primitives. Note that this result finally describes the time-varying scene structures modeled by a set of saliency primitives. Regardless of the noisy inputs explained above, saliency primitives allow us to deal with underlying primitive patterns in the extracted spatiotemporal patterns since the identification includes the estimation of noise variance. Since the primitives contain translations, deformations (resizes) and saliency variations of salient regions, they are capable of describing visual events caused by distinct objects.

The OSDM presented here will be utilized in Chapter 4 to take into account of time-varying scene structures when describing the spatiotemporal correlations.

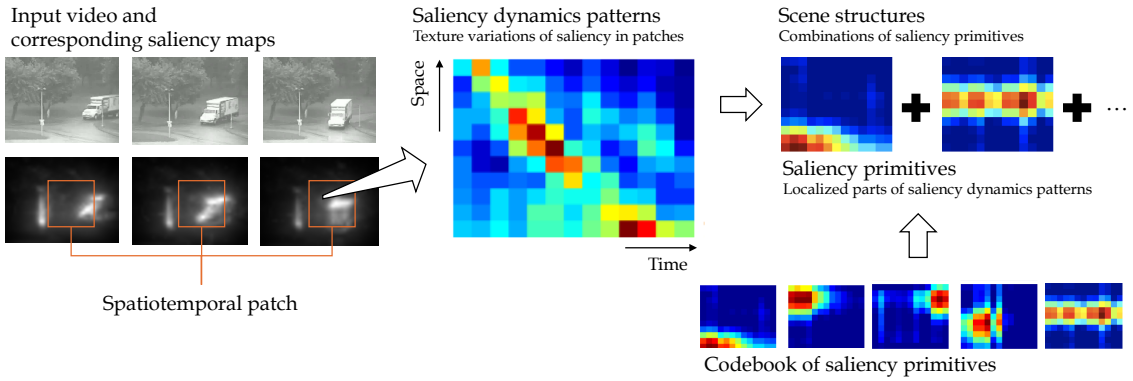

Figure 2.7: Overview of the patch-based saliency dynamics model. Parts of the images in this figure are contained in the dataset provided by [Mahadevan and Vasconcelos, 2010].

2.4 Patch-Based Saliency Dynamics Model

2.4.1 Overview of the Model

Next, we present the patch-based saliency dynamics model, which takes a great advantage when dealing with unedited natural videos such as surveillance videos. The basic options are listed as follows (see also Figure 2.7):

1. We apply a sliding-window approach to define a temporal interval and try to describe the saliency dynamics in each interval. In other words, we do not particularly focus on scene change events.

2. We regard temporal variations in the textures of saliency maps in spatiotemporal patches as saliency dynamics patterns and try to utilize them for modeling scene structures consisting of visual events caused by not only distinct objects but also textures (like natural sceneries and human crowds).

3. The saliency primitives are utilized to describe those texture variations of saliency in a direct manner (i.e., sequences of multivariate vectors). It allows us to involve various changes including complex variations jointly caused by more than one salient regions.

Along with the OSDM, we assume that the saliency dynamics patterns are independent for each temporal interval.

Although the direct representation adopted in saliency primitives allows us to deal with complex variations caused by a variety of visual events, we need

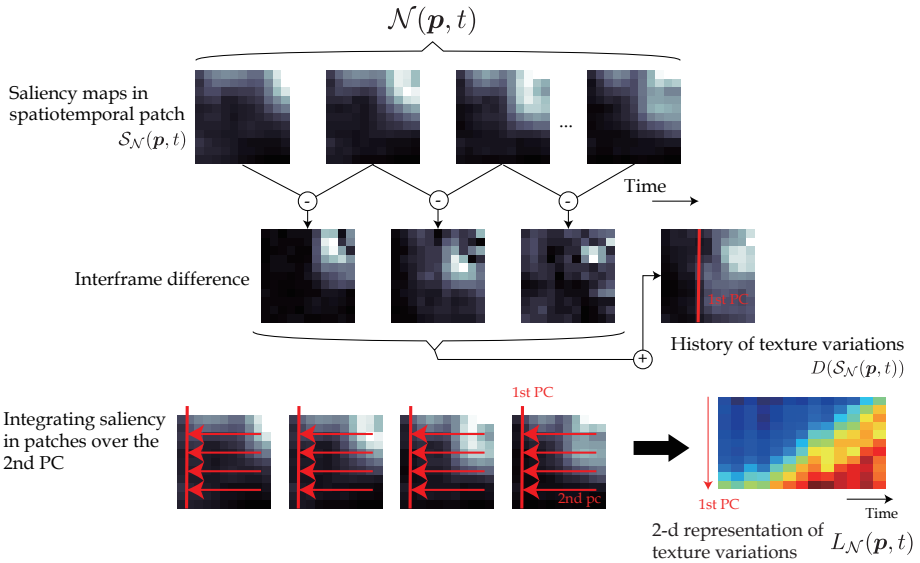

Figure 2.8: Extracting texture variations of saliency maps.

an efficient and robust modeling so as to cope with the diversity and noise in the saliency dynamics as discussed in Section 2.2.2. To this end, we introduce a codebook of saliency primitives, where the primitives describe localized parts of saliency dynamics patterns in a direct manner, like in the right of Figure 2.7. By statistically learning the codebook from videos so that each primitive describes the parts frequently appearing the videos, we can achieve the efficiency as well as the robustness when describing saliency dynamics patterns.

In the following sections, We first present a method to extract texture variations of saliency maps in a spatiotemporal patch as saliency dynamics patterns (Section 2.4.2) and then learning method of the codebook (Section 2.4.3).

2.4.2 Extracting Texture Variations of Saliency Maps

Let us denote a spatiotemporal patch around $(\boldsymbol{p},t)=(x,y,t)$ as

$$\begin{aligned}\mathcal{N}(\boldsymbol{p},t) &:= \Omega_{(\delta_x,\delta_y)}\times\mathcal{T}_{\delta_t},\\ \Omega_{(\delta_x,\delta_y)} &\subseteq [x-\delta_x,x+\delta_x]\times[y-\delta_y,y+\delta_y],\ \mathcal{T}_{\delta_t}\subseteq[t-\delta_t,t+\delta_t],\end{aligned}\quad (2.3)$$

where $\delta_x,\delta_y,\delta_t$ define the size of patch. Although we can essentially define δ_x and δ_y independently, in what follows we use the same size $\delta_x=\delta_y=\delta_s$ and denote the spatial patch as Ω_{δ_s} for simplicity. Then, a spatiotemporal volume of saliency

maps cropped by the patch is denoted as follows:

$$S_{\mathcal{N}(\boldsymbol{p},t)} = \left(S_{(\Omega_{\delta_\mathrm{s}},\min(\mathcal{T}_{\delta_\mathrm{t}}))},\ldots,S_{(\Omega_{\delta_\mathrm{s}},\max(\mathcal{T}_{\delta_\mathrm{t}}))}\right).$$

$S_{\mathcal{N}(\boldsymbol{p},t)}$ contains the texture variations of saliency maps in a spatiotemporal patch, which is regarded as saliency dynamics patterns in this model. If $\Omega_{\delta_\mathrm{s}} = \Omega$, $S_{\mathcal{N}(\boldsymbol{p},t)}$ leads to the description of overall scene structures like the OSDM. Otherwise, i.e., $\Omega_{\delta_\mathrm{s}} \subset \Omega$, we can describe a kind of local scene structures in a given patch.

To avoid diversity in the saliency dynamics patterns, we particularly focus on their amplitude when extracting the texture variations. In other word, we introduce an orientation-invariant description for the extracted texture variations. Specifically, we first look for an axis in a spatial domain to describe the amplitude of texture variations the best (see also Figure 2.8). We calculate the absolute inter-frame differences in $\mathcal{S}_{\mathcal{N}(\boldsymbol{p},t)}$ and sum them up over time to obtain the history of the texture variation, $f_{\mathcal{N}(\boldsymbol{p},t)} : \Omega_{\delta_\mathrm{s}} \to \mathbb{R}_+$ of the following form:

$$f_{\mathcal{N}(\boldsymbol{p},t)}(\boldsymbol{p}) = \sum_{\tau\in\mathcal{T}_{\delta_\mathrm{t}}} \left| S_{(\Omega_{\delta_\mathrm{s}},\tau)}(\boldsymbol{p}) - S_{(\Omega_{\delta_\mathrm{s}},\tau-1)}(\boldsymbol{p}) \right|.$$

We then approximate $f_{\mathcal{N}(\boldsymbol{p},t)}$ by massive samples and apply the principal component analysis to the samples to obtain two principal component axes $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ in the spatial domain, where the first component $\boldsymbol{u}_1$ describes an orientation of the maximum variation of the history.

Finally, we sum up the degrees of saliency over $\boldsymbol{u}_2$ for every frame to get the 2-d representation of the texture variations, $\|u_1\| \times (2\delta_\mathrm{t}+1)$ matrix $L_{\mathcal{N}(\boldsymbol{p},t)} = (L_{\min(\mathcal{T}_{\delta_\mathrm{t}})},\ldots,L_{\max(\mathcal{T}_{\delta_\mathrm{t}})})$ where $L_\tau = \sum_{u_2} S_{(\Omega_{\delta_\mathrm{s}},\tau)}$ (we will define $\|\boldsymbol{u}_1\|$ later). Practically, we can obtain L_τ by rotating an image describing $S_{(\Omega_{\delta_\mathrm{s}},\tau)}$ so that $\boldsymbol{u}_1$ corresponds to the horizontal direction in a new 2-d coordinate and summing up the rotated image the over vertical direction in the coordinate. Thus, the spatial size of the variation, $\|\boldsymbol{u}_1\|$, is given as the maximum length of the line segment with slope $\tan(\boldsymbol{u}_1)$. That is, $\|\boldsymbol{u}_1\|$ satisfies $2\delta_\mathrm{s}+1 \le \|\boldsymbol{u}_1\| \le \sqrt{2}(2\delta_\mathrm{s}+1)$. Since size $\|\boldsymbol{u}_1\|$ has a variation among samples according to the angle of $\boldsymbol{u}_1$, we crop $L_{\mathcal{N}(\boldsymbol{p},t)}$ so that each sample has the same size (practically $2\delta_\mathrm{s}+1$) in what follows.

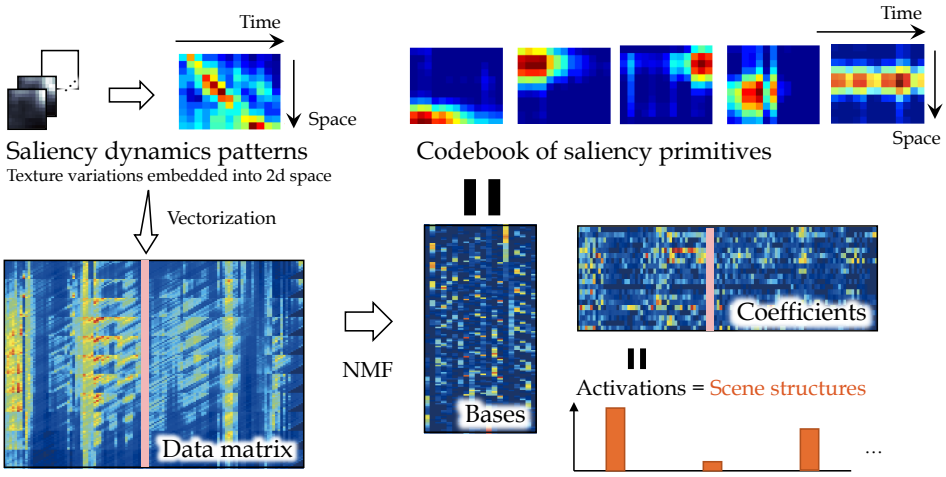

Figure 2.9: Learning a codebook of saliency primitives.

2.4.3 Learning a Codebook of Saliency Primitives

Given many samples of saliency dynamics patterns (texture variations of saliency) extracted in the above procedure, we learn a codebook consisting of saliency primitives that describe localized parts of the patterns. Since the saliency dynamics patterns characterize scene structures consisting of multiple visual events, they can contain the mixture of several dynamics. For this reason, a standard model of dynamic textures that introduces a single LDS for each spatiotemporal patch such as [Doretto et al., 2003] is not always appropriate for our situations. Instead, mixture models such as [Chan and Vasconcelos, 2008, Ravichandran et al., 2009] can describe such dynamics patterns with a set of submodels. The PSDM that utilizes saliency primitives to describe parts of the dynamics patterns can be regarded as the latter approach. In what follows, we aim to learn a codebook of saliency primitives effectively via matrix factorization.

Let us denote a flatten vector of saliency dynamics pattern $L_{\mathcal{N}(\boldsymbol{p},t)}$ as $\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p},t)} \in \mathbb{R}_+^J$ where $J = (2\delta_s + 1) \cdot (2\delta_t + 1)$. We introduce a codebook consisting of N saliency primitives, $\mathcal{D} = \{D_1, \ldots, D_N\}$, where $D_n \in \mathbb{R}_+^J$ is the flatten vector of primitive patterns defined in the same spatiotemporal domain as $\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p})}$. Then, $\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p},t)}$ can be described with $\boldsymbol{w}(\boldsymbol{p},t) = (w_1, \ldots, w_N)^{\mathrm{T}} \in \mathbb{R}_+^N$, where w_n is the degree of activation for primitive D_n (i.e, how strongly primitive D_n appears). By setting $N < J$, we can introduce an efficient description for $\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p},t)}$.

To learn codebook $\mathcal{D}$, we adopt a non-negative matrix factorization (NMF) [Lee and Seung, 1999] (see Figure 2.9). NMF plays an effective role in face analysis [Lee and Seung, 1999], music transcription [Smaragdis and Brown, 2003], doc-

ument clustering [Xu et al., 2003], etc. It decomposes a non-negative matrix into two non-negative factors, where one factor consists of localized and structured bases and the other has corresponding sparse activation coefficients like Figure 2.9. Let us introduce a $J \times N_{\mathrm{sp}}$ data matrix containing N_{sp} samples of saliency dynamics patterns, $\mathcal{L} = (\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p}_1, t_1)}, \ldots, \boldsymbol{l}_{\mathcal{N}(\boldsymbol{p}_{N_{\mathrm{sp}}}, t_{N_{\mathrm{sp}}})})$. Then, NMF derives the two factors as follows:

$$\mathcal{L} = \bar{\mathcal{D}} W + \mathcal{E},$$

where $\bar{\mathcal{D}} = (D_1, \ldots, D_N)$, a $J \times N$ basis matrix, represents a sequence of saliency primitives (that is, the codebook $\mathcal{D}$), $W = (\boldsymbol{w}(\boldsymbol{p}_1, t_1), \ldots, \boldsymbol{w}(\boldsymbol{p}_{N_{\mathrm{sp}}}, t_{N_{\mathrm{sp}}}))$, an $N \times N_{\mathrm{sp}}$ coefficient matrix, is corresponding activations, and $\mathcal{E}$ is a residual. $\bar{\mathcal{D}}$ and W can be obtained based on the optimization as follows:

$$\min_{\bar{\mathcal{D}}, W} \frac{1}{2} \|\mathcal{L} - \bar{\mathcal{D}} W\|_{\mathrm{F}}^2, \text{ s.t., } \bar{\mathcal{D}}, \ W \geq 0.$$

We solve the above optimization problem by adopting multiplicative update rules [Lee and Seung, 2001] implemented in [Li and Ngom, 2013].

2.4.4 Examples

This section introduces examples of model estimation results with a public dataset including unedited natural videos. Specifically, we employed ASCMN database [Riche et al., 2012][(iii)], which contained 24 videos consisting of outdoor scenes, surveillance videos, videos of human crowds, etc. We adopted the Itti's saliency map [Itti et al., 1998], where the features include the luminance, color and orientation. We particularly focused on the local scene structures as a unique product of the PSDM compared to the OSDM, and investigated several sizes of patches: $(\delta_{\mathrm{x}}, \delta_{\mathrm{y}}, \delta_{\mathrm{t}}) = (5 \text{ pixel}, 5 \text{ pixel}, 0.4 \text{ sec})$, and $(15 \text{ pixel}, 15 \text{ pixel}, 0.4 \text{ sec})$ [(iv)] , where the videos were first resized into 80×60 pixel resolution. Note that the spatial sizes of patches were 11×11 pixel and 31×31 pixel in the above settings. When learning the codebook of saliency primitives, we preliminarily resized 31×31 pixel patches into 11×11 pixel. In the following examples, the size of codebook N was empirically set to $N = 20$.

Figure 2.10 depicts selected examples of extracted saliency dynamics patterns as well as corresponding videos and saliency maps. These patterns were extracted

[(iii)] `http://www.tcts.fpms.ac.be/attention/?categorie13/databases`

[(iv)] Since the frame rate of videos was 15 fps, the interval 0.4 sec was regarded as 6 frames.

at the point where a single subject looked at, and describing local scene structures in a spatiotemporal patch. The points of gaze, the red points in the 2nd column of the figure, are located at the center point of the dynamics patterns in the 4th column due to the definition of $\mathcal{N}(\boldsymbol{p}, t)$ in Equation (2.3). These examples demonstrate that the points of gaze are not always directed to the most salient locations in a spatiotemporal patch, such as 5th and 6th rows in Figure 2.10. In other words, there are sometimes spatiotemporal gaps between saliency and gaze dynamics. The yellow points in the 2nd row of Figure 2.10 describe gaze scan-paths around the red gaze points, which indicate the large gaze motions can provide large spatiotemporal gaps. In this way, the local scene structures modeled by the PSDM can contribute to the analyses of the event-level spatiotemporal gaps in spatiotemporal correlations. We will revisit these phenomena in Chapter 5.

Figure 2.11 shows the comparison between extracted dynamics patterns and reconstructed ones from learned primitives as well as the degrees of activations for each pattern. As discussed in Section 2.3.5, the variations of saliency cannot always be continuous since saliency maps are obtained in a frame-wise manner. However, several discontinuities were smoothed in the reconstructed patterns as shown in the right of Figure 2.11. It indicates that the proposed model is capable of obtaining brief patterns while avoiding noises.

Finally, Figure 2.12 shows the samples of codebooks obtained from 11×11 pixel patches and 31×31 pixel patches. These codebooks contribute to the efficient description of scene structures since the number of primitives in the codebooks ($N = 20$) are much smaller than the original sizes of patches ($J = (2 \cdot 6 + 1) \cdot (2 \cdot 5 + 1) = 143$). Thanks to the modeling by the NMF, the obtained primitives successfully describe localized parts. As shown in the figure, the primitives sometimes contain several salient regions. Although the increases in the sizes of codebooks can suppress these phenomena, the sizes should be determined based on the objectives of gaze behavior analyses. That is, we can introduce a cross-validation scheme to determine the size based on the performance of a certain gaze behavior analysis.

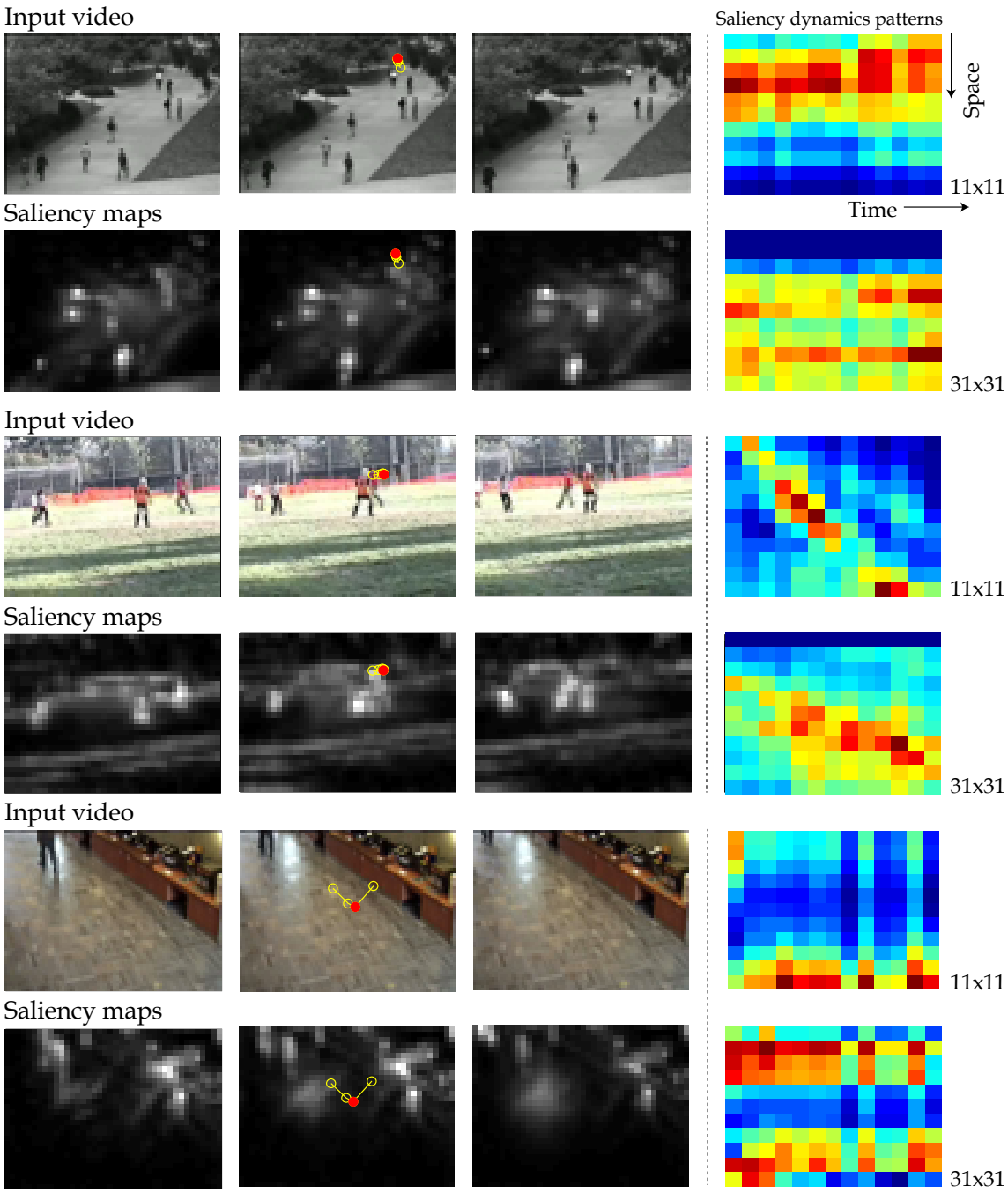

Figure 2.10: Examples of saliency dynamics patterns in spatiotemporal patches of different sizes. The patterns in the 4th column are extracted at gaze points of a single subject, which is denoted as the red points in the 2nd column of input images and saliency maps. Parts of the images in this figure are contained in the dataset provided by [Mahadevan and Vasconcelos, 2010, Itti and Baldi, 2009, Li et al., 2004].

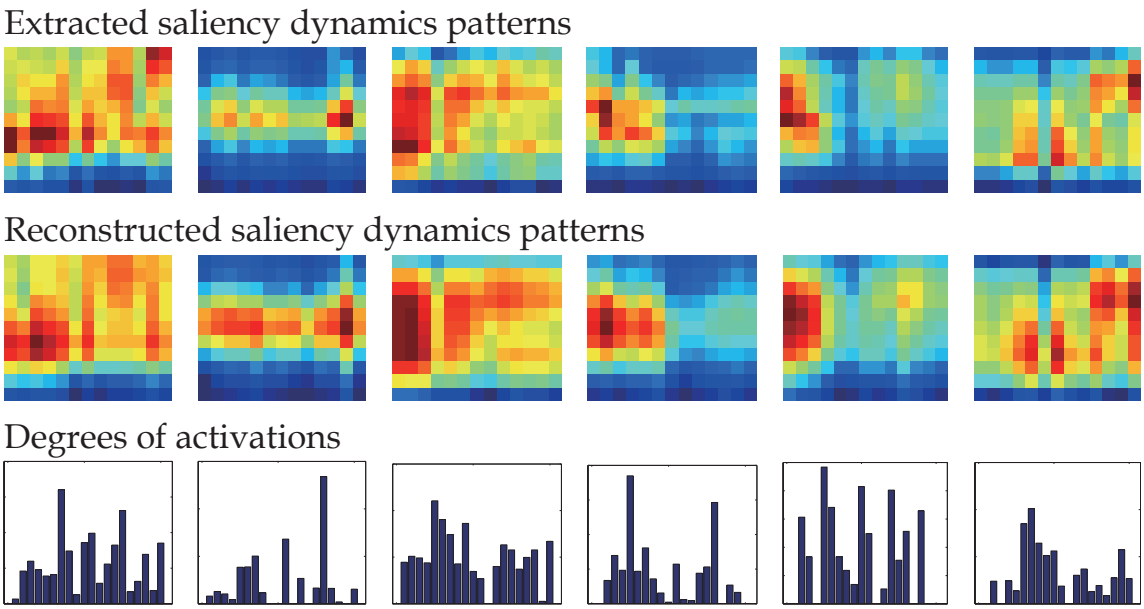

Figure 2.11: Extracted saliency dynamics patterns, reconstructed patterns and the degrees of activations.

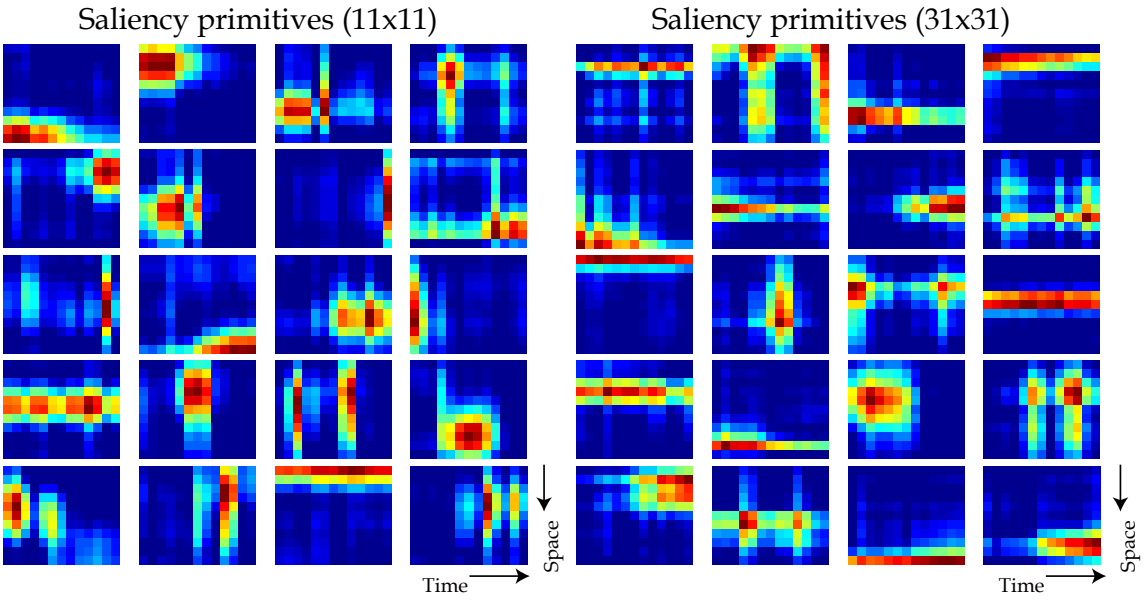

Figure 2.12: Examples of codebooks.

Chapter 3

Attentional Target Identification Using Temporal Synchronizations

3.1 Introduction

This chapter and following two chapters are aimed at assessing the effectiveness of our framework by describing spatiotemporal correlations and evaluating them via practical gaze behavior analyses in real environments. Although the saliency dynamics models introduced in the Chapter 2 allow us to handle various visual events and time-varying scene structures using saliency primitives, this chapter adopts manually-designed videos with a constant scene structure and saliency primitives given preliminarily. Thanks to this simplification, we can concentrate on the evaluation of spatiotemporal correlations.

We particularly address event-level spatiotemporal gaps in the spatiotemporal correlations and investigate how they appear in actual gaze behavior. Imagine the situations where we are browsing dynamic contents with visual events like Figure 3.1. In the example, three items generate visual events (object translations) in a certain temporal interval. When we examine one of them (the center one in the example), a reaction to the translations will appear in our gaze dynamics almost at the same time. This is a temporal synchronization between visual events and gaze reactions, and we aim to describe it with the event-level spatiotemporal gaps in our framework. Specifically, suppose first that the spatiotemporal patterns caused by the visual events are represented by saliency primitives and the patterns of primitives as well as the exact times that the primitives appear are given. Then, we detect gaze primitives corresponding to the reactions by matching the template reflecting the patterns of primitives. Finally, we can calculate

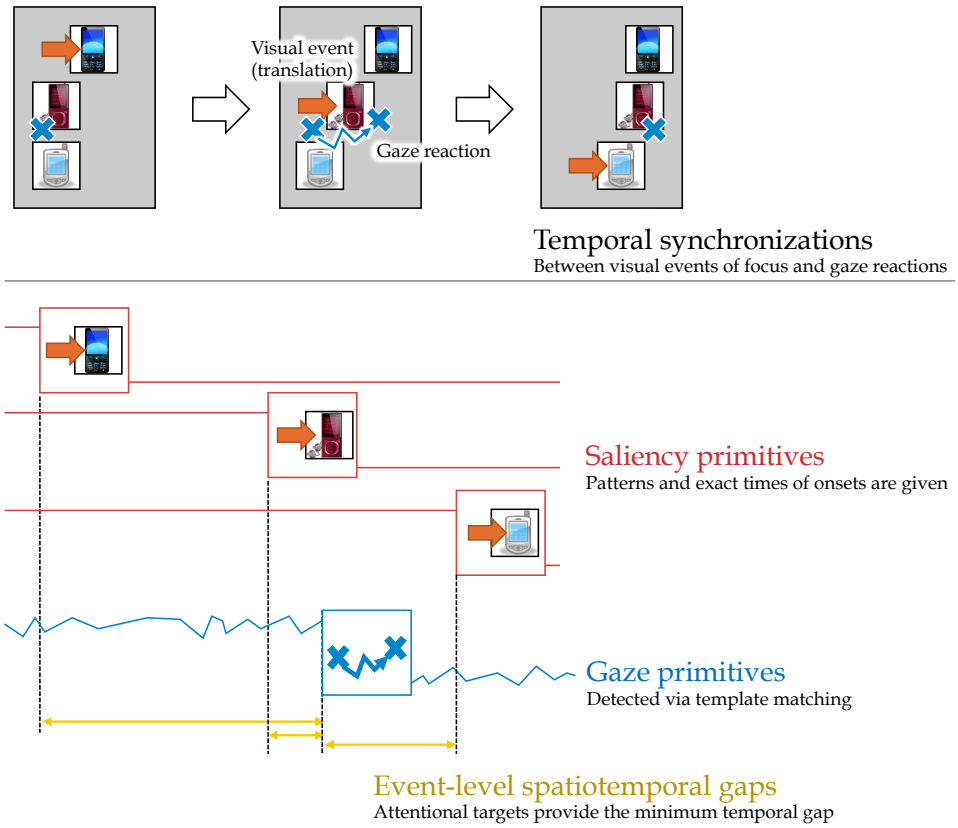

Figure 3.1: Describing event-level spatiotemporal gaps. The temporal distances between saliency and gaze primitives represent the temporal synchronizations between visual events and gaze reactions.

temporal distances between the onset times of saliency and gaze primitives as a descriptor of temporal synchronizations (see Figure 3.1).

We leverage this synchronization modeled as event-level spatiotemporal gaps for the task of identifying attentional targets from visual contents with several distinct objects (attentional target identification) to evaluate their effectiveness. The attentional target identification is an important task for designing dynamic contents such as navigation systems and advertisements in our daily life. Intuitively, the most naive approach is to use the spatial locational relationships between the objects and the points of gaze; given regions of objects, we can identify targets by judging which object regions is the closest to the points of gaze. However, this approach is not always effective when gaze tracking systems involve a large measurement error. Generally, gaze tracking techniques have the trade-off between their accuracy and the allowable range of subjects' poses and positions. For example, active tracking techniques using corneal

reflections (e.g., [Zhu and Ji, 2005, Hennessey et al., 2006, Chen et al., 2008], Review [Morimoto and Mimica, 2005]) perform high accuracy while restricting the subjects' poses and positions. On the other hand, appearance-based approaches, e.g., [Beymer and Flickner, 2003, Ishikawa et al., 2004, Wang et al., 2005], often allow subjects to change their poses and positions instead of the accuracy.

In this study, we propose a identification method based on the temporal synchronizations, which we refer to as the *Gaze Probing*. The Gaze Probing regards the objects with saliency primitives that provide the minimum temporal distances to reactions as attentional targets. Since gaze tracking errors affect the only template matching to detect gaze primitives of reactions, we can achieve a robust identification by designing saliency primitives and templates appropriately. The following sections first introduce specific details as to the Gaze Probing including some discussions on how to design saliency primitives and how to detect gaze primitives from gaze data. Then, we evaluate the performance of the Gaze Probing based on experiments that subjects freely browsed several designed contents.

3.2 The Gaze Probing

Assume the situations where human observers are browsing multiple objects in dynamic contents to choose one of them. The contents are supposed to display a constant scene structure with given types of visual events. For examples, Figure 3.1 depicts an example of the dynamic content consisting of three image objects with translation events.

In this section, we first introduce a description of event-level spatiotemporal gaps between saliency primitives and corresponding gaze primitives for measuring temporal synchronizations. Then, we discuss how to design saliency primitives (i.e., what types of visual events are acceptable) and how to detect gaze primitives for the attentional target identification by the Gaze Probing.

3.2.1 Describing Event-level Spatiotemporal Gaps

Let us denote a set of objects in dynamic contents as $\{O_c | c = 1, ..., C\}$. These objects are supposed to be distinguished from each other so as to be easily tracked by observers, while they are possibly overlapped to each other or out of frame temporarily. We denote properties of the c-th object region as $\boldsymbol{\theta}_t^{(c)} \in \mathbb{R}_+^J$ and their spatiotemporal pattern as $\Theta^{(c)} = (\boldsymbol{\theta}_1^{(c)}, \ldots, \boldsymbol{\theta}_T^{(c)})$.

This chapter does not adopt saliency dynamics models to extract and identify saliency primitives from contents and instead manually design and embed the primitives into object motions. We denote designed primitive D in a direct form, such as $D = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{\delta_\mathrm{t}})$, where the δ_t is the size of the primitive. Then, we embed multiple instances of D in $\Theta^{(c)}$, where the i-th onset of primitives is located at $t_i^{(c)}$. That is, $\Theta^{(c)}$ is partially defined as follows:

$$\boldsymbol{\theta}_t^{(c)} = \boldsymbol{d}_{t-t_i^{(c)}+1} \quad \left(t_i^{(c)} \leq t \leq t_i^{(c)} + \delta_\mathrm{t} - 1\right). \tag{3.1}$$

Note that the remaining parts of $\Theta^{(c)}$ can be interpolated arbitrarily so as not to obtain more saliency than the primitives.

The Gaze Probing measures temporal synchronizations between visual events and gaze reactions as the event-level spatiotemporal gaps (specifically, temporal distances) between the onsets of designed primitives and those of gaze primitives detected from gaze data. To investigate the temporal synchronizations clearly, we design overall scene structures so that all the primitives embedded in multiple objects must have temporally different onsets to each other with an enough margin. Namely, for arbitrary pairs of objects $O_c, O_{c'}$ $(c \neq c')$ and pairs of IDs i and i', $t_i^{(c)}, t_i^{(c')}, t_{i'}^{(c)}$ and $t_{i'}^{(c')}$ must satisfy $\left|t_i^{(c)} - t_{i'}^{(c)}\right| \geq \varepsilon$ and $\left|t_i^{(c)} - t_i^{(c')}\right| \geq \varepsilon$, where the minimum margin, ε, should be large enough to distinguish it from a reaction delay. Such scene structures allows us to discriminate the designed primitives in synchronization from those provided by the others.

Once we detect the onset of gaze primitives corresponding to gaze reactions at frame T_react, we can calculate the temporal distances between the onsets as event-level spatiotemporal gaps. Specifically, we introduce an evaluation score for each instance of designed primitives such as $V_i^{(c)} = |T_\mathrm{react} - t_i^{(c)}|$.

3.2.2 Gaze Probing for Attentional Target Identification

Identification and interpolation

In the Gaze Probing, a target, $O_{\hat{c}}$, can be identified as follows based on the evaluation score introduced in the preceding arguments:

$$\hat{c} = \arg\min_c V_i^{(c)}.$$

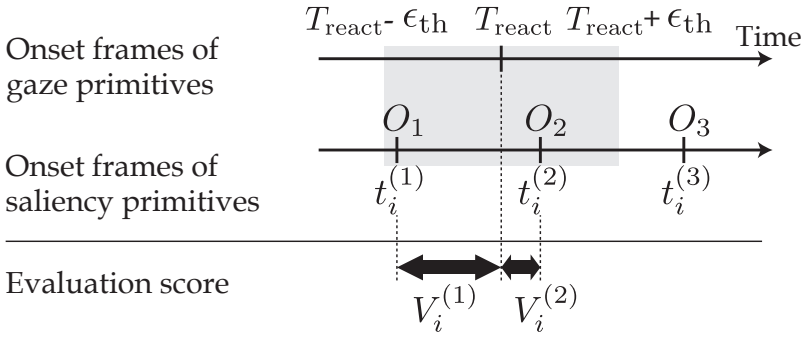

Figure 3.2: Measuring temporal synchronizations. We consider saliency primitives only in the gray region, $[T_{\text{react}} - \epsilon_{\text{th}}, T_{\text{react}} + \epsilon_{\text{th}}]$.

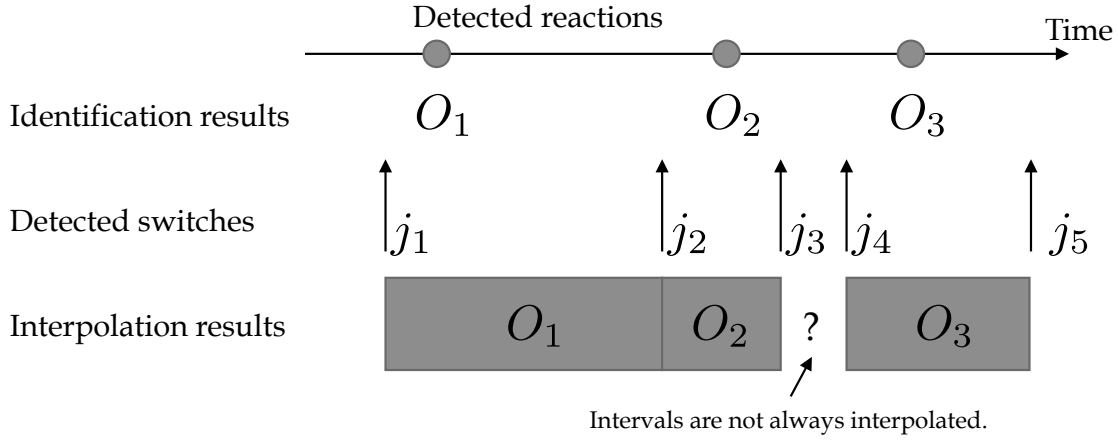

Figure 3.3: Interpolation of the identification results.

Practically, we set threshold ϵ_{th} to $V_i^{(c)}$ in order to avoid irrelevant synchronizations (see Figure 3.2). Namely, if $V_i^{(c)}$ is larger than the threshold, we regard the corresponding reaction as false positive detection.

Since the Gaze Probing deals with event-level relationships, identification results provided above is intermittent. If we additionally detect the switches of targets from gaze data, we can partially interpolate the results. Let us assume that a set of switches is detected at frames $\{j_1, \ldots, j_{N_{\text{int}}}\}$ (see also Figure 3.3). We first define fixation intervals between successive switches, such as $[j_1, j_2), \ldots, [j_{N_{\text{int}}-1}, j_{N_{\text{int}}}]$. For each interval, we can extend the identification results if reactions take place in the interval; if a reaction is detected at $j_{n-1} \leq T_{\text{react}} \leq j_n$ and is associated with $t_i^{(\hat{c})}$, attentional targets within $[j_{n-1}, j_n]$ are identified as $O_{\hat{c}}$.

Designing saliency primitives

Attentional target identification based on the Gaze Probing basically needs to observe temporal synchronizations between visual events and gaze reactions clearly. The design of scene structures provided in Section 3.2.1 is indeed one of the techniques to address this issue. In addition, there are several requirements for the design of saliency primitives (in other words, the types of visual events acceptable in the Gaze Probing), in particular if we make the identification robust to gaze tracking errors while allowing human observers to behave freely.

Basically, the primitives need to involve a distinct spatiotemporal pattern (1) that can be well reflected in gaze dynamics, and (2) that enables us to detect the onset of gaze primitives of reactions easily. We therefore set practical requirements for saliency primitives as follows.

(1) Requirements to reflect saliency primitive patterns in gaze dynamics

(1a) Short-term patterns. Saliency primitives should be reflected in gaze dynamics even when observers switch attentional targets frequently and pursue each targets in a short term. That is, the temporal size of saliency primitives, δ_t, is required to be shorter than fixation durations against targets.

(1b) Simple and slow patterns. As mentioned in the preceding section, the temporal distances between instances of primitives, ε, should be larger than a reaction delay. In other words, saliency primitives should contain a pattern providing small delays. We can expect the small delays if the pattern is simple and slow motions enough to pursue.

(2) Requirements to detect the onset of gaze primitives accurately

(2a) Distinguishable from gaze tracking errors. Gaze reactions become indistinguishable from gaze tracking errors if the amplitudes of spatiotemporal patterns are extremely small. Thus, the patterns are required to be larger than an average gaze tracking error at least.

(2b) Distinguishable from endogenous target examinations. Gaze behavior involves not only exogenous motions that occur when pursuing targets in motion but also endogenous actions to browse the targets. The amplitude of saliency primitives is supposed to be larger than the size of objects in order to distinguish the endogenous ones from the exogenous ones.

(2c) Distinguishable from endogenous target switches. Endogenous actions can also occur when observers switch targets. Thus, the direction of object motions in saliency primitives should be orthogonalized to that of constituent objects to distinguish them. This also helps the interpolation of identification results presented in the preceding section.

Considering the requirements presented so far, we adopt an *onset of horizontal scrolls* as a simple saliency primitive as illustrated in Figure 3.4 (1). We first assume saliency primitives defined in Equation (3.1) only involve horizontal translations (i.e., $\boldsymbol{\theta}_t^{(c)}, \boldsymbol{d}_t \in \mathbb{R}_+$ denotes horizontal locations of objects). The primitives embedded as the i-th primitive in the c-th object are denoted as follows:

$$\boldsymbol{d}_t = \begin{cases} b_i^{(c)} & (t_i^{(c)} \leq t < t_i^{(c)} + \tau_{\text{onset}}) \\ (t - t_i^{(c)} - \tau_{\text{onset}}) m_i^{(c)} + b_i^{(c)} & (t_i^{(c)} + \tau_{\text{onset}} \leq t \leq t_i^{(c)} + \delta_{\text{t}}), \end{cases} \tag{3.2}$$

where τ_{onset} is the frame of the onset. Namely, the c-th object that is embedded i-th primitive remains stationary at $b_i^{(c)}$ from $t_i^{(c)}$ to $t_i^{(c)} + \tau_{\text{onset}}$, and starts scrolling with velocity $m_i^{(c)}$ until frame $t_i^{(c)} + \delta_{\text{t}}$.

Since the contents of interests are assumed to displayed in real environments, we need to avoid unnatural designs of content dynamics. For that purpose, we suppose that object motions are similar and highly correlated to each other. Specifically, we equalize the velocity of all the scrolls so that the saliency of all the objects' motion are as equal as possible (i.e., $m_i^{(c)} = m$ for any i and c). On that basis, we first set δ_{t} small enough to meet with Requirement (1a) (the actual value will be shown in the experiment section). The speed of scrolls, m, is supposed to be smaller than the maximum speed of humans' pursuit eye movements, 40°/sec [Vision Society of Japan, 2000], for Requirement (1b). In addition, the amplitude of scrolls, $(\delta_{\text{t}} - \tau_{\text{onset}})m$, is supposed to be larger than the size of the measurement error as well as that of objects due to Requirements (2a) and (2b). Finally, we limit the direction of scroll to be horizontal and line up objects vertically so as to cope with Requirement (2c).

Detecting Reaction Primitives from Gaze Data

This section presents a method to detect gaze primitives corresponding to reactions from gaze data. Let us denote a sequence of horizontal gaze positions as $X = (x_1, \ldots, x_T)$. Ideally, the spatiotemporal pattern of saliency primitives, D, appears in X as is, and we can detect the onset of reactions, T_{react}, via the template

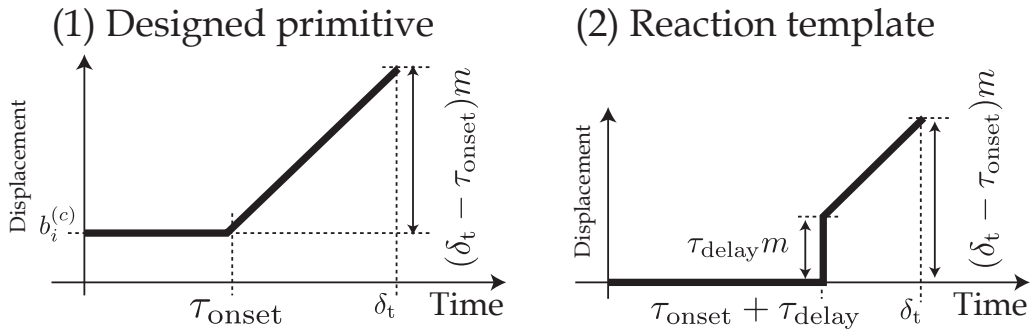

Figure 3.4: Designed saliency primitive and the corresponding reaction template.

matching of D in X. However, gaze reactions toward targets in motion often involve a reaction delay as shown in [Rashbass, 1961, Joiner and Shelhamer, 2006] and thus do not always contain the same patterns as D.

We therefore introduce a template considering the reaction delay as shown in Figure 3.4 (2). Specifically, we define a modified template primitive, $D' = (\boldsymbol{d}'_1, \ldots, \boldsymbol{d}'_{\delta_{\text{t}}})$ as follows:

$$\boldsymbol{d}'_t = \begin{cases} 0 & 0 \leq t < \tau_{\text{onset}} + \tau_{\text{delay}}) \\ (t - \tau_{\text{onset}})m & (\tau_{\text{onset}} + \tau_{\text{delay}} \leq t \leq \delta_{\text{t}}), \end{cases}$$

where τ_{delay} is the size of a reaction delay defined preliminarily. Humans sometimes initiate saccades before pursuing objects in motion so as to capture the objects in their central fovea with less retinal blurs. Template Θ'' considers this characteristic by the adding a reaction delay and a sudden motion from position 0 to $\tau_{\text{delay}}m$. We use template D' to detect onsets of gaze primitives, T_{react}. Specifically, we calculate the normalized cross correlation (ZNCC) between D' and parts of X extracted by a sliding window of size δ_{t} to detect T_{react} as the frame where the ZNCC haves local maxima with the value larger than threshold c_{th}. Practically, c_{th} is set to be slightly smaller than 1 in order to allow the gaze tracking errors while avoiding incorrect detections of reactions.

Note that the design of saliency primitives and detection of gaze primitives in this chapter contribute to the robustness to gaze tracking errors obviously, since the template matching needs not use vertical gaze locations which contain larger errors than horizontal in many cases [Zhu and Ji, 2005, Chen et al., 2008, Wang et al., 2005].

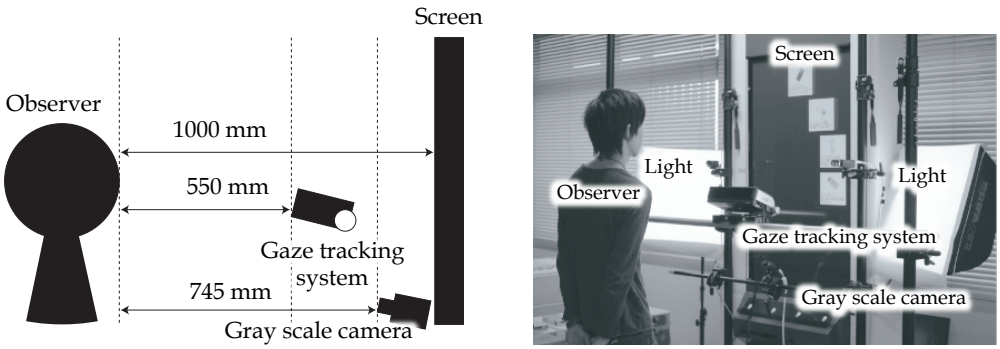

Figure 3.5: Displaying and sensing system with a large screen.

3.3 Experiments

In the experiments, we evaluated how much event-level spatiotemporal gaps can describe gaze behavior of subjects freely browsing the contents via the task of attentional target identification. Specifically, we implemented practical dynamic contents that followed the requirements shown in the previous section, and then compared the proposed Gaze Probing with several other baselines.

3.3.1 System Setups

We built a display system with a large screen[(i)] shown in Figure 3.5. The distance between subjects and the system was approximately 1000 mm. Under such settings, subjects were able to look at a whole area of the screen, and at the same time they were supposed to move their eyes large enough to measure. We placed a gray-scale camera[(ii)] to capture the frontal face of subjects all the time during experiments. Since the camera was not capable of zooming, panning and tilting, the subjects were only allowed to change their head poses and positions as long as their eyes were captured by the camera. We additionally placed two lights[(iii)] to get a light intensity enough to detect irises.

We adopted an appearance-based approach to the gaze tracking. Specifically, we tracked facial feature points from videos based on the Active Appearance

[(i)]FUJITSU UBWALL. The size of screen is 1106 mm in height and 622 mm in width. The height at the center of the screen is 1462 mm. We used upper 562 mm areas for content playback.

[(ii)]Point Grey Research Grasshopper (1600×1200 pixel, 8 bit gray scale, 30 fps, 1/1.8 inch CCD), with FUJINON HF16HA-1B (f=16 mm).

[(iii)]RIFA-F (500×500 mm).

Model [Cootes et al., 2001], fitted a 3-d face model and detected iris positions to obtain a gaze point sequence. The obtained sequence was calibrated with the given positions of markers on a screen and corresponding positions of gaze points when staring at the markers; we obtained an affine transformation individually to horizontal and vertical directions to calibrate the obtained sequence. The average sizes of gaze tracking errors on the screen were 62 mm (3.6°) horizontally and 94mm (5.4°) vertically.

In addition to the appearance-based tracking presented so far, we also adopted an IR-based gaze tracking system[(iv)] for a ground-truth label of gaze points. While the average size of gaze tracking errors was 27 mm (1.6°), this system limited head poses and positions of subjects in the range of IR lights, which was narrower than that of the appearance-based tracking. We thus constrained the behavior of subjects so as to use the IR-based method stably.

3.3.2 Evaluation Scheme

Let us denote the number of gaze primitives detected from gaze data as r_{detected} and that of correct identifications of attentional targets as r_{success}. Then, identification precision R_{prec} is defined as $R_{\text{prec}} = r_{\text{success}}/r_{\text{detected}}$. We also evaluated how much designed primitives were reflected in gaze data. Specifically, we counted the number of primitives r_{all} that subjects actually looked at with a ground-truth data to calculate the recall of the identification, $R_{\text{reca}} = r_{\text{success}}/r_{\text{all}}$.

We additionally employed several baselines as follows:

Position-based method ($\mathbf{M}_{\text{pos}}$) The position-based method focuses on the spatial locational relationships. Assume that $\boldsymbol{\theta}_t^{(c)}, \boldsymbol{p}_t \in \mathbb{R}_+^2$ respectively describe the 2-d (horizontal and vertical) locations of the c-th object and gaze points for this case. Then, we calculate pairwise distances between objects and gaze points at each frame and identify attentional targets as follows:

$$\hat{c} = \arg\min_c \left\| \boldsymbol{\theta}_t^{(c)} - \boldsymbol{p}_t \right\|. \tag{3.3}$$

Since this method obtains results in a frame-wise manner, we evaluate precision score R_{prec} with the number of frames and that of correct identifications. In addition, recall R_{reca} is the same as R_{prec} in this method because we involve all the frames into the identification.

[(iv)] Tobii X120 Eye Tracker, 60 Hz, The allowable range of head motion is 400×220×300 mm.

Correlation-based method ($\mathbf{M}_{\text{corr}}$) This method calculates the ZNCC between horizontal object motions and gaze data in interval $[\min_c t_i^{(c)}, \max_c t_i^{(c)} + \delta_{\text{t}}]$ to identify the targets as the ones performing the highest correlation. Specifically, the target is identified based on the following criteria:

$$\hat{c} = \arg\max_c \frac{\sum_{\min t_i^{(c)}}^{\max t_i^{(c)}+\delta_{\text{t}}} \left(\theta_t^{(c)} - \bar{\theta}^{(c)}\right)(x_t - \bar{x})}{\sqrt{\sum_{\min t_i^{(c)}}^{\max t_i^{(c)}+\delta_{\text{t}}} \left(\theta_t^{(c)} - \bar{\theta}^{(c)}\right)^2} \sqrt{\sum_{\min t_i^{(c)}}^{\max t_i^{(c)}+\delta_{\text{t}}} (x_t - \bar{x})^2}}, \tag{3.4}$$

where $\bar{\theta}^{(c)}$ and $\bar{x}$ are the average value of the sequence. The precision is obtained as $R_{\text{prec}} = r_{\text{success}}/r_{\text{detected}}$. Since this baseline evaluates all the intervals that contain designed primitives, R_{reca} is the same as R_{prec}.

Hybrid method ($\mathbf{M}_{\text{squ}}$) This method considers both the spatial distance and the motion correlation by the sum of absolute differences as follows ($\boldsymbol{\theta}_t^{(c)}, \boldsymbol{p}_t \in \mathbb{R}_+^2$ respectively describe the 2-d locations along with the position-based method):

$$\hat{c} = \arg\min_c \sum_{\min_c t_i^{(c)}}^{\max_c t_i^{(c)}+\delta_{\text{t}}} \left\| \boldsymbol{\theta}_t^{(c)} - \boldsymbol{p}_t \right\|^2. \tag{3.5}$$

The scores are obtained as $R_{\text{prec}} = R_{\text{reca}} = r_{\text{success}}/r_{\text{detected}}$.

3.3.3 Experiments with Artificial Contents

Design of experiments

We first conducted experiments with relatively artificial settings, which utilized small plain-color rectangles as objects. Since the objects contain no information to examine, we can expect subjects conduct less endogenous actions.

Six subjects were individually asked to look at one of the objects for 20 sec $\times$ 2 sessions, where they were able to switch targets during the session. Two objects colored with gray (20 mm$\times$ 20 mm[(v)]) served as the objects and they performed reciprocating scrolling motions from the left to right of the screen. The amplitude of the scroll was 400 mm and the speed, m, was constantly 466 mm/sec (25.0°). The interval of each cycle was set to 4 sec and the objects made a short stop at each

[(v)]Biologically, the size of central fovea is approximately 2° [Wright and Ward, 2008] that corresponds to 20 mm in the screen from the viewpoint of subjects.

Table 3.1: Precision and recall scores for artificial contents by the Gaze Probing M_{prop} and the comparative methods M_{pos}, M_{corr}, and M_{squ}.

	M_{pos}	M_{corr}	M_{squ}	M_{prop} (R_{prec})	M_{prop} (R_{reca})
Score [%]	64.6	74.8	69.1	**85.3**	83.8

edge of the scroll. We set reaction delay τ_{delay} to $\tau_{delay} = 0.15$ sec by taking psychophysical findings [Rashbass, 1961, Joiner and Shelhamer, 2006] into account. On that basis, the onsets of designed primitives in different objects have a constant temporal gap, 0.4 sec, which is larger than τ_{delay}. The size of the designed primitives, δ_t, was empirically set to 1.0 sec. We placed the two objects vertically, where the distance between them was set to 150 mm so as not to be smaller than gaze tracking errors.

In the proposed Gaze Probing, threshold ϵ_{th} to avoid irrelevant associations between designed primitives and reactions was empirically set to 0.5 sec. In addition, the threshold to detect the reaction primitives, c_{th}, was set to 0.9 to allow gaze tracking errors while avoiding incorrect detections of reactions.

Results and discussions

Table 3.1 introduces the quantitative results of each method, and Figure 3.6 depicts some illustrative examples of gaze data. The number of gaze primitives being detected, $r_{detected}$, was 109 in the total 240 sec, which was 98.2 % of all the primitives being looked at.

The Gaze Probing $\mathbf{M}_{prop}$ obtained the highest score compared with the other baselines. Since the distance between objects were comparatively close to the size of gaze tracking errors in the vertical direction, the scores of position-based method M_{pos} decreases. In addition, object motions are highly correlated with each other and thus all the objects often have higher scores in correlation-based and hybrid methods, M_{corr}, M_{squ}, which seemed to result in the decrease of their scores. On the other hand, the Gaze Probing considers the only temporal information and thus performs the robustness to gaze tracking errors as well as the similarity of object motion patterns.

We also evaluated the performance of the interpolation. Switches of targets were detected as the local maxima in the vertical acceleration of gaze dynamics. The interpolated frames that were applied the identification results were 94.7 %,

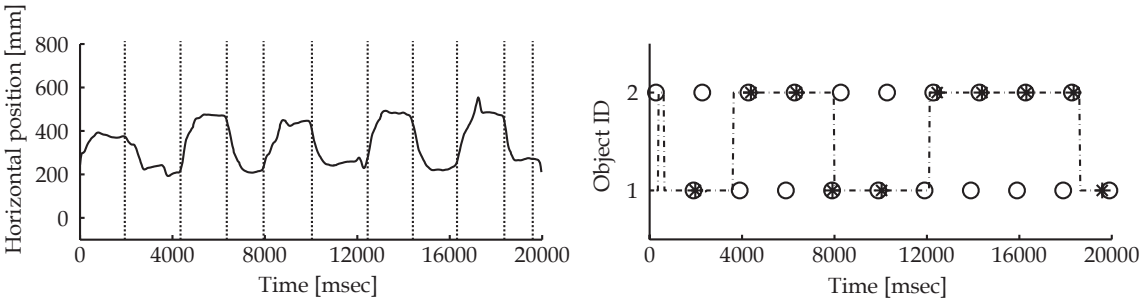

Figure 3.6: Examples of gaze data and identification results. Left: gaze data (solid line) and reactions (dot line), Right: designed primitives (o), reactions (*) and the verified object (dashed line). This figure is a part of author's publication [Yonetani et al., 2010] copyrighted by Human Interface Society Japan.

where the precision in a frame-wise manner was 84.5 % in average. Although the switches were not always detected correctly due to the vertical tracking errors, we were able to interpolate most of the frames and the precision there was still higher than that by baseline methods.

3.3.4 Experiments with Natural Contents

Design of experiments

We now address a more natural situation that photos and text captions are displayed as object items in a catalog content. In particular, the object items consisted of a photo of cellular phones and a text caption below the photo about the specifications of items (about 50 Japanese characters). The size of the objects were set to 150 mm$\times$150 mm. The six subjects were asked to choose the mosts interesting items from the displayed objects for 60 sec. In these situations, the subjects were expected to browse the content freely with not only with exogenous motions but possibly with endogenous actions such as examining and comparing the items.

Specifically, we adopted the following two designs for the contents.

[D-1] Reciprocal swinging design (Figure 3.7) This design involves objects swinging horizontally with a natural motion. First, each object scrolled to the left. As an object approached the left edge of the screen, the object smoothly slowed until it stopped at the edge for a short period of time. The object then scrolled to the right in a similar manner. Saliency primitives in this case are defined as

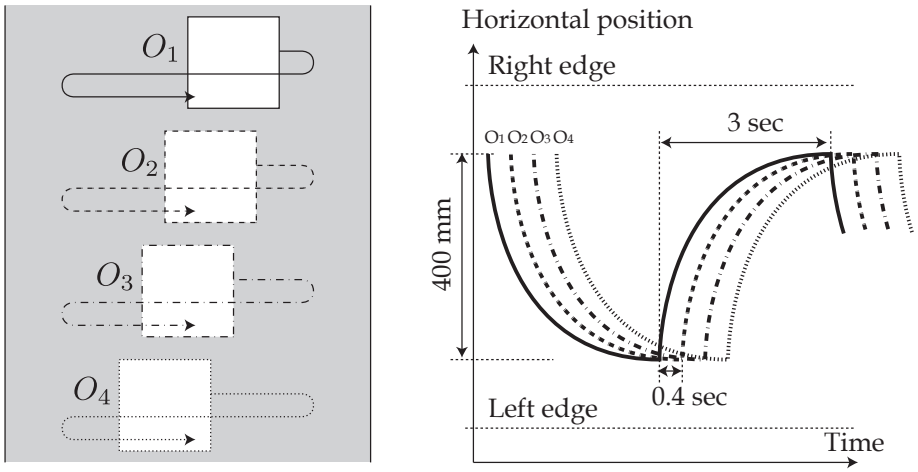

Figure 3.7: [D-1] Reciprocal swinging design. This figure is a part of author's publication [Yonetani et al., 2010] copyrighted by Human Interface Society Japan.

the scrolling motions from stopping at each edge. The temporal interval between the primitives for a given object was set to 3 sec while the period between the primitives for two different objects was set to 0.4 sec.

[D-2] Scrolling motion design (Figure 3.8) This design assumes the situations that items are dynamically updated over time to provide as much information as possible to observers. As an object appeared at the right of the screen approached to the left, it slowed down with a smooth motion. The object stopped at the left edge shortly and finally disappeared from the screen. Following this, the object displaying *different items* appeared at the right edge once again, and moved in the same manner. There were three items displayed for each object; these items were updated at each new appearance of the objects. Saliency primitives in this case are defined as the scrolling motions from stopping at the right edge. The temporal interval between the primitives for a given object was set to 6 sec while the period between the primitives of two objects was set to 0.4 sec.

Parameter settings As common settings for the two designs, they have four objects aligned vertically with 10 mm gaps as shown in Figure 3.7 and Figure 3.8. While we adopted the onset of horizontal scrolls as designed primitives, the objects stopped with a smooth slowdown for the sake of naturalness.

Unlike the artificial settings in Section 3.3.3, short-term gaze motions can appear frequently to switch and examine attentional targets. We considered Re-

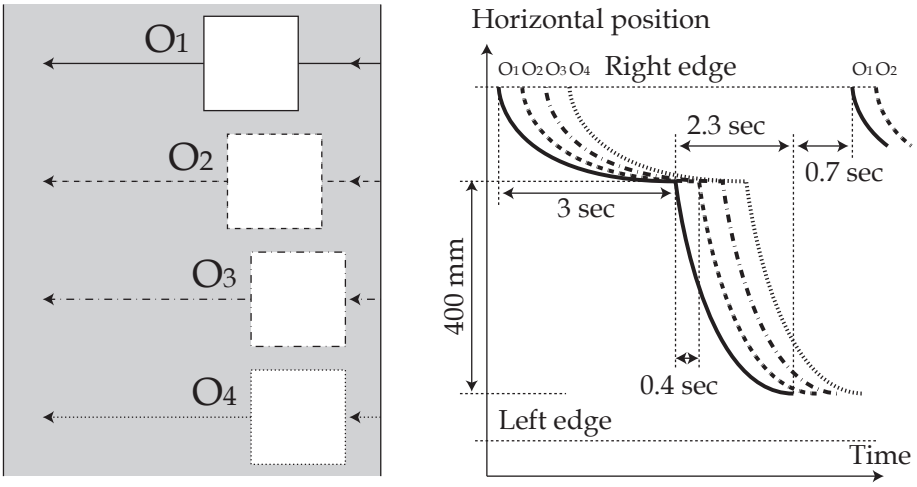

Figure 3.8: [D-2] Scrolling design. This figure is a part of author's publication [Yonetani et al., 2010] copyrighted by Human Interface Society Japan.

quirement (2b) in Section 3.2.2 and set $(\delta_t - \tau_{onset})m$ so as to be larger than not only gaze tracking errors but gaze motions while the examination. Specifically, we set $(\delta_t - \tau_{onset}) = 0.64$ sec (then, $(\delta_t - \tau_{onset})m = 300$ mm) and $\tau_{onset} = 0.6$ sec so that we can judge if subjects stare at the objects. Note that all the objects remain stationary longer than τ_{onset} to make subjects examine images and texts to some extent; the objects stopped for 1.43sec in [D-1] and 1.63sec in [D-2] while considering naturalness.

In the proposed method, thresholds ϵ_{th} and c_{th} were set to 0.5 sec and 0.9, respectively, the same as the previous settings in Section 3.3.3.

Results and discussions

Table 3.2 shows quantitative results and Figure 3.9 depicts some samples of gaze data and reactions. The number of gaze primitives being detected was 98 times (90.7 % of all the primitives being looked at) in [D-1] and 56 times (93.3 %) in [D-2] for the overall 360 sec.

Although the Gaze Probing obtained the highest precision score, it sometimes fails accurate detections of gaze primitives, which resulted in the decrease of overall scores particularly in [D-1]. Comparing [D-1] with [D-2], we found more endogenous actions like target examinations and switches around the onsets of saliency primitives in [D-1] than in [D-2], which leaded to the above decrease. On the other hand, subjects seemed to attract objects in [D-2] in an exogenous manner since items displayed in the objects changed over time in [D-2].

Table 3.2: Precision and recall scores for designed contents by the proposed method M_{prop} and the comparative methods M_{pos}, M_{corr}, and M_{squ}. [D-1]: swinging design, [D-2]: scrolling design.

[D-1]	M_{pos}	M_{corr}	M_{squ}	M_{prop} (R_{prec})	M_{prop} (R_{reca})
Score [%]	54.5	60.1	63.3	**68.4**	62.0
[D-2]	M_{pos}	M_{corr}	M_{squ}	M_{prop} (R_{prec})	M_{prop} (R_{reca})
Score [%]	41.9	45.7	51.4	**76.8**	71.7

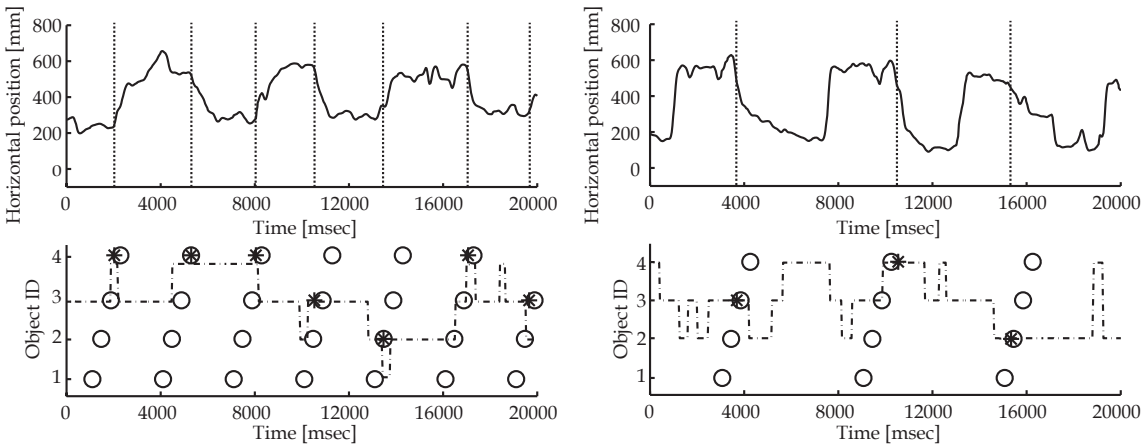

Figure 3.9: Examples of gaze data and identification results for [D-1] (left) and [D-2] (right). Above: gaze data (solid line) and reactions (dot line), below: designed primitives (o), reactions (*) and the verified object (dashed line). This figure is a part of author's publication [Yonetani et al., 2010] copyrighted by Human Interface Society Japan.

We also evaluated the effectiveness of interpolation. We took the same approaches to the detection of the switches of targets as the previous experiments in Section 3.3.3. The intervals that can obtain identification results by the interpolation was 74.0% for [D-1] and 45.4% for [D2], while the precision in those intervals was 63.4% and 61.6%, respectively. Since gaze behavior toward [D-2] contained less endogenous actions as discussed in the previous argument, the interpolatable intervals become shorter in that design. One possible solution to this problem is to embed a large number of designed primitives as long as preserving the naturalness of the contents.

3.4 General Discussions

3.4.1 Gaze Tracking Errors and Identification Accuracies

Since any approach to the attentional target identification analyzes gaze dynamics, the performance of identification depends on gaze tracking errors. Methods M_{pos} and M_{squ} involve the evaluation of spatial distances, $|\boldsymbol{\theta}_t^{(c)} - \boldsymbol{p}_t|$ as shown in Equation (3.3) and Equation (3.5). Thus, their performance decreases when gaze tracking systems provide a bias error for the direction of constituent objects (i.e., vertical in our experiments). M_{corr} is robust to the bias error since x_t is normalized as shown in Equation (3.4). However, it requires to observe gaze primitives in gaze data accurately, and thus it decreases the performance if the gaze data contain spontaneous random errors for the direction of the spatiotemporal patterns of primitives (i.e., horizontal in our experiments). Compared with the methods discussed above, the Gaze Probing is basically affected by the tracking errors only when detecting gaze primitives via template matching. While the special template considering a reaction delay contributes to higher precision scores, it can be affected by the spontaneous random errors along with M_{corr}.

To evaluate the characteristics of the Gaze Probing in more detail, we added a Gaussian noise to the ground-truth data in Section 3.3.3 and evaluated how precision scores can change over the size of errors. We considered the mean and standard deviation (SD) of errors in the appearance-based method (94 mm in vertical and 45 mm in horizontal directions, respectively) and set the mean and SD of the Gaussian noise as 70$\sim$110 mm and 25$\sim$65 mm, respectively. Figure 3.10 provides the variation of accuracies by all the methods over the changes of the mean and SD of the noises. As can be seen in the above of the figure, the Gaze Probing and M_{corr} are not affected by the size of bias errors. In addition, the bottom of the figure shows that the Gaze Probing works more stably than M_{corr} as long as the SD of the noises is approximately the same as that of the appearance-based gaze tracking adopted in the experiments, while the precision scores decrease according to the size of the SD. Table 3.3 describes the variations in the ratio of gaze primitives correctly detected with the variations of the SD. Although the ratio decreases as the SD becomes large, it keeps 92$\sim$98% and does not seem to affect the final scores so much. Consequently, we can successfully observe gaze primitives by achieving Requirements (1a) (1b), although sometimes fail to achieve (2a) to detect them accurately due to the random noises.

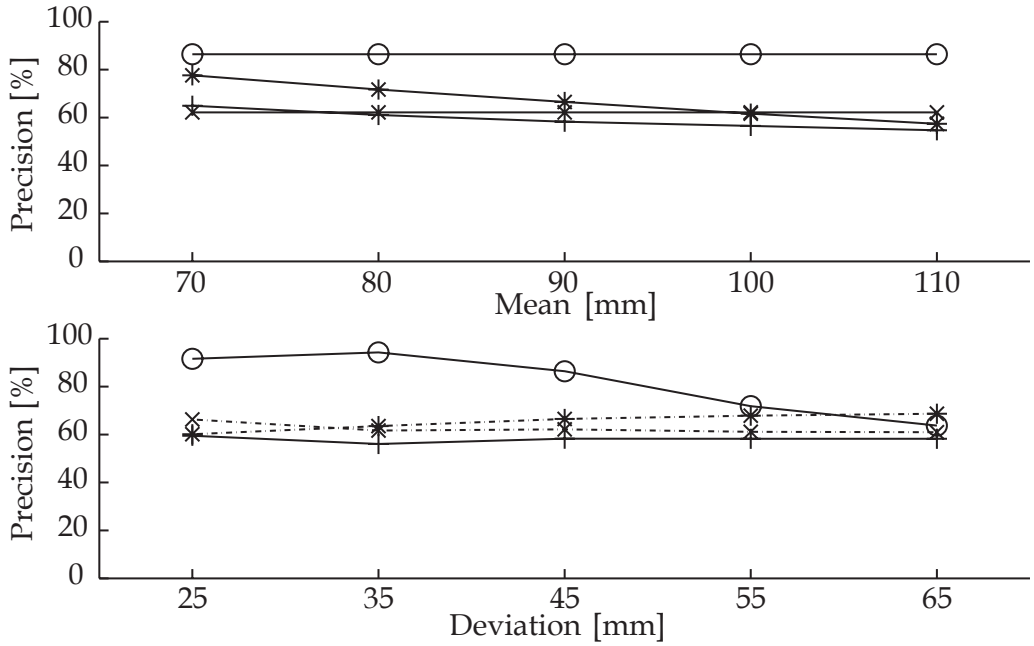

Figure 3.10: Precision scores under various different settings of noises. Above: scores at five different means (SD=90mm), below: scores at five different SDs (mean=90mm). Legends indicate M_{pos}(*), M_{corr}(x), M_{squ}(+), M_{prop}(o). This figure is a part of author's publication [Yonetani et al., 2010] copyrighted by Human Interface Society Japan.

As a result of the discussion so far, it is important for the Gaze Probing not only to design saliency primitives accurately but also to introduce an effective approach to the detection of gaze primitives. In this study, we adopted a simple template matching for the detection. Since the matching technique is aware of the local changes of gaze dynamics, it is likely to be affected by random noises in gaze tracking systems. To address this problem, we can involve global information of gaze dynamics via scale-space analyses [Witkin, 1983], for example. Tracking the onset of reactions from course to fine can perform a more accurate detection of gaze primitives while avoiding random noises.

3.4.2 Designing Dynamic Contents

In practical cases, the aim of dynamic contents is to attract humans' attention, especially when the contents is displayed by navigation systems, recommender systems and so on. In such applications, our design of the dynamic contents has the potential to give unnatural impressions to observers. In other words, the

Table 3.3: The ratio of reaction primitives correctly detected at five different SDs (mean=90mm).

SD [mm]	25	35	45	55	65
Ratio [%]	98.2	94.6	92.3	92.3	93.7

Gaze Probing still has room for improvement of saliency primitive designs. We conducted a brief interview as to the impression to the content designs employed in Section 3.3.4. The results revealed that subjects tended to feel unnaturalness or stress on [D-1] than [D-2]. This seems to be because designed of contents in [D-1] apparently seems to have no intentions while those in [D-2] is aimed at updating items being displayed. To adopt the Gaze Probing in practical cases, one of the future work is to investigate the naturalness of saliency primitives in detail.

Chapter 4

Attentive State Estimation based on Video Scene Structures

4.1 Introduction

This chapter is aimed at assessing our framework under a situation where human observers are watching intentionally-designed videos such as TV commercial films. While the designed contents adopted in Chapter 3 contained a constant scene structure with given types of visual events (i.e., translations of objects), the videos that we will address in this chapter involve time-varying scene structures due to various types of visual events including deformations and saliency variations of objects as well as frequent scene changes. Therefore, we now introduce the proposed framework with the object-based saliency dynamics model (OSDM) and try to handle visual events and scene structures with help from saliency primitives achieved by fitting the OSDM to videos.

Within the framework, we particularly focus on scene-level correlations between scene structures and gaze dynamics, which is the other aspect of spatiotemporal correlations that was not addressed in the previous chapter. The aim of this chapter is to describe how the scene-level correlations can be characterized differently depending on the time-varying types of scene structures (see Figure 4.1). To this end, we first classify saliency primitives and gaze primitives into several types based on their spatiotemporal patterns. Then, the types of scene structures can be featured by the combinations of saliency primitive types. In addition, we leverage the classified type information for features that describe scene-level correlations effectively as follows:

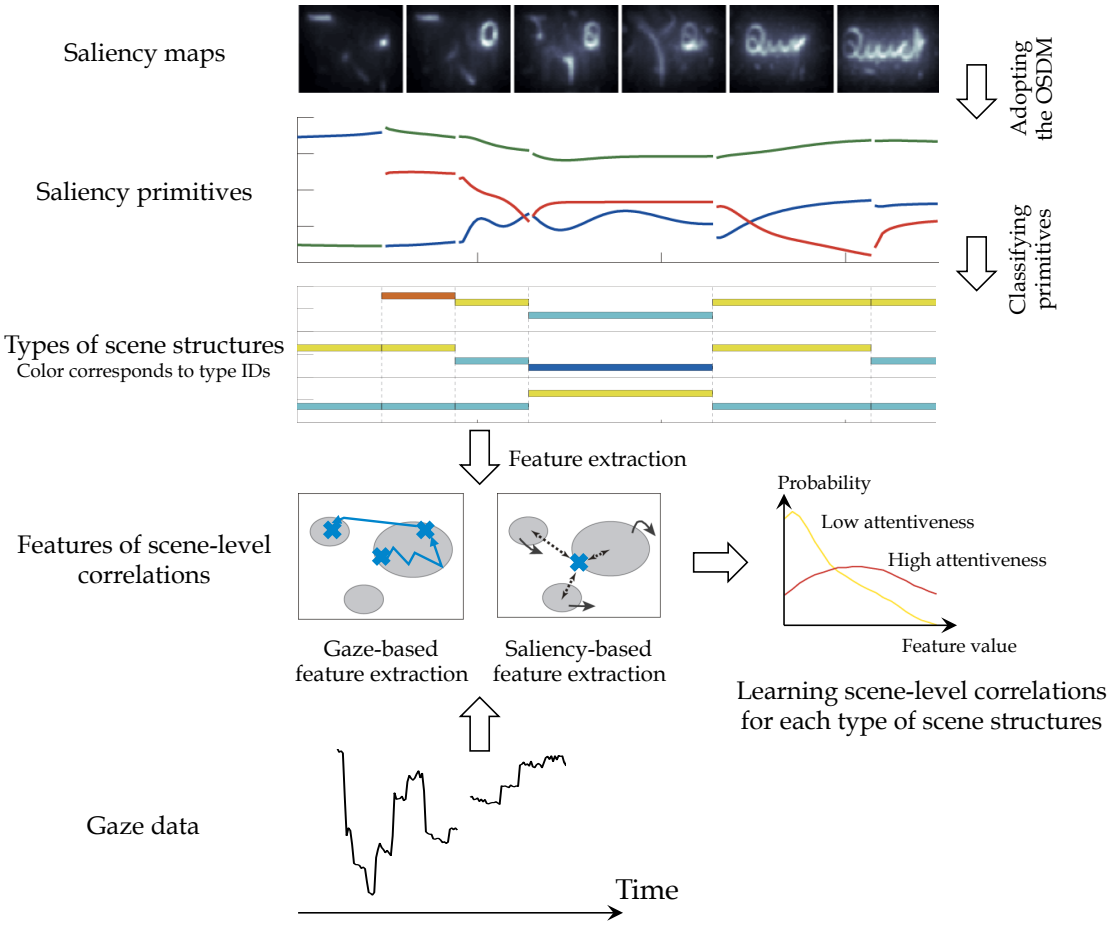

Figure 4.1: Describing scene-level correlations using the OSDM.

Gaze-based feature extraction focuses on how specific types of gaze primitives can be characterized when looking at a certain type of saliency primitives.

Saliency-based feature extraction examines which types of saliency primitives originally tend to be looked at in a certain type of scene structures.

As for a task of gaze behavior analyses, this chapter employs attentive state estimation that classifies if human observers concentrate on displayed videos or not. This task is a kind of mental state estimation reviewed in Section 1.2.2, which is now a popular problem in the fields of HCI and visual psychology. In particular, the estimation of attentive states has many applications including interactive system designs. Specifically, the levels of attentiveness serve as a crucial clue for giving a feedback from the systems in a timely manner.

The proposed descriptions of scene-level correlations influenced by the types of scene structures are effectively utilized for the attentive state estimation as be-

low. First, we train discriminative models of attentive states with features of scene correlations for each type of scene structures. Then, given a new pair of video and gaze data, we adaptively apply the trained models based on the identified types of scene structures. It enables us to estimate attentive states when watching videos while considering time-varying scene structures.

In the following sections, we first present the basic formulations of the proposed attentive state estimation in Section 4.2. Then, the twofold feature extraction schemes including the classification methods of saliency and gaze primitives will be introduced in Section 4.3 and Section 4.4.

4.2 Attentive State Estimation

4.2.1 Formulation

As for the basis of our attentive state estimation, we follow a traditional approach to mental state estimation based on a supervised learning framework like [Bednarik et al., 2012, Eivazi and Bednarik, 2011]. It begins with extraction of features from gaze data such as frequencies of saccades and durations of fixations. At the same time, feature samples in a training dataset are given one of the several labels indicating discrete mental states (in this study, the level of attentiveness). Then, the mental state estimation is formulated as a problem of learning a discriminative model for these labels. Practically, existing studies adopt the support vector machine (SVM) [Bednarik et al., 2012], hidden Markov models (HMM) [Eivazi and Bednarik, 2011] and so on.

Let us introduce gaze data $X = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T)$. We denote feature vectors extracted from X as $\boldsymbol{\varphi}(\boldsymbol{X}) \in \mathbb{R}^{N_{\text{feat}}}$, where N_{feat} is the number of the features. At the same time, we consider discrete labels $A \in \{A_1, \ldots, A_{N_{\text{state}}}\}$, where N_{state} is the number of mental states to be considered. Then, the estimation can be formulated as a standard classification problem based on the posterior probability of A with observation $\boldsymbol{\varphi}(\boldsymbol{X})$:

$$\hat{A} = \arg\max_{A} P\left(A \mid \boldsymbol{\varphi}(X)\right). \tag{4.1}$$

Based on the formulation in Equation (4.1), the proposed method involves time-varying scene structures as well as scene-level correlations derived by fitting the OSDM introduced in Section 2.3 (see Figure 4.1 for detail). Let us assume that X is split into $(X_1 \ldots, X_K)$ based on scene segmentation $\mathcal{I} = (I_1, \ldots, I_K)$.

Each interval I_k has a set of spatiotemporal patterns of salient regions, $\mathbf{\Theta}_k = \{\Theta_k^{(1)}, \ldots, \Theta_k^{(C_k)}\}$, where $\mathbf{\Theta}_k^{(c)}$ is modeled by saliency primitive $D_k^{(c)}$. We classify $D_k^{(c)}$ into several types and describe the types of scene structures by the combinations of types identified to the primitives. Specifically, let us consider a set of possible saliency primitive types $\mathcal{W} = \{w_1, \ldots, w_N\}$ where N is the number of types. Given a scene structure modeled by a set saliency primitives in the k-th interval, $\mathcal{D}_k = \{D_k^{(1)}, \ldots, D_k^{(C_k)}\}$, we first classify $D_k^{(c)}$ into one of several types, which is denoted as $W_k^{(c)} \in \mathcal{W}$. Then, the type of scene structures at the k-th interval is modeled as a vector consisting of histogram counts of $W_k = \{W_k^{(1)}, \ldots, W_k^{(C_k)}\}$, which is denoted as $hist(W_k)$. Finally, we use scene structure $\mathcal{D}_k$ and its type $hist(W_k)$ to modify Equation (4.1) as follows:

$$\hat{A}_k = \arg\max_A P\left(A \mid \boldsymbol{\varphi}(X_k, \mathcal{D}_k), hist(W_k)\right), \tag{4.2}$$

where $\boldsymbol{\varphi}(X_k, \mathcal{D}_k) \in \mathbb{R}^{N_{\text{feat}}}$ is a feature vector describing the scene-level correlations between gaze dynamics X_k and scene structures $\mathcal{D}_k$. This formulation describes an adaptive estimation of attentive states based on time-varying types of scene structures, $hist(W_k)$. If we can obtain feature vectors in a frame-wise manner for each $\boldsymbol{p}_t \in X_k$, Equation (4.2) can be rewritten as follows:

$$\hat{A}_t = \arg\max_A P\left(A \mid \boldsymbol{\varphi}(\boldsymbol{p}_t, \mathcal{D}_k), hist(W_k)\right). \tag{4.3}$$

As briefly introduced in the previous section, the gaze-based feature extraction utilizes gaze primitives for their description. That means it requires gaze-point sequence X_k to identify the types of gaze primitives and thus follows the formulation of Equation (4.2). On the other hand, the saliency-based extraction does not consider such gaze dynamics particularly and thus it can conduct the estimation in a frame-wise manner based on Equation (4.3).

4.2.2 Feature Extraction from Saliency Primitives

Classification of saliency primitives is important for describing not only the types of scene structures but also extracting features of scene-level correlations in the proposed method. Although we will introduce different classification techniques for each feature extraction in Section 4.3 and Section 4.4, we here introduce a common feature extracted from saliency primitives adopted to their classification.

As presented in Section 2.3.3, saliency primitive $D_k^{(c)}$ in the OSDM is derived

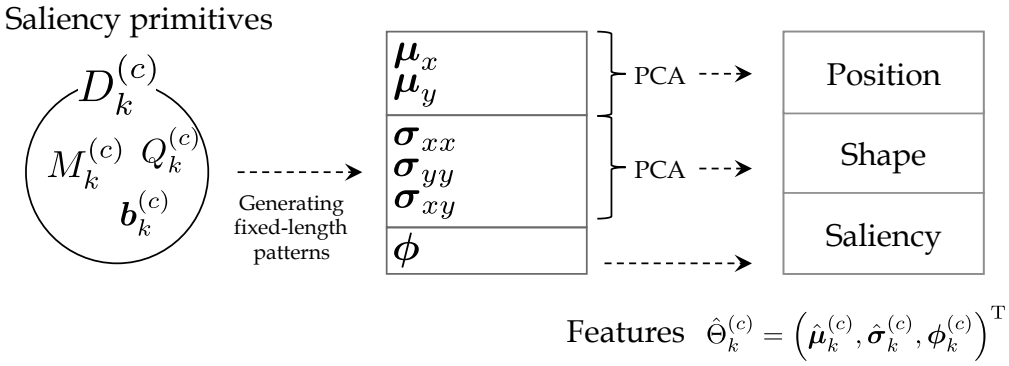

Figure 4.2: Feature extraction from saliency primitives for their classification.

from the six-dimensional spatiotemporal patterns consisting of mean vectors, variance and covariance elements and a weight factor of the GMM, where they correspond to positions, shapes and the degree of saliency, respectively. In this study, we particularly focus on the amplitudes of motions, resizes and the variations of saliency as the characteristics of saliency primitives and sacrifice the orientations of motions and resizes for simplicity (see Figure 4.2). Specifically, we first generate saliency dynamics pattern $\boldsymbol{\Theta}_k^{(c)}$ of the constant size T_{gen} from saliency primitive $D_k^{(c)}$. In what follows, we omit subscriptions c and k without loss of generality. Then, $\boldsymbol{\Theta}$ can be described by the concatenation of six column vectors such as $\boldsymbol{\Theta} = \Big((\boldsymbol{\mu}_x, \boldsymbol{\mu}_y), (\boldsymbol{\sigma}_{xx}, \boldsymbol{\sigma}_{yy}, \boldsymbol{\sigma}_{xy}), \boldsymbol{\phi}\Big)$, where $(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)$ correspond to horizontal and vertical means, $(\boldsymbol{\sigma}_{xx}, \boldsymbol{\sigma}_{yy}, \boldsymbol{\sigma}_{xy})$ for two variances and one covariance, and $\boldsymbol{\phi}$ for a weight factor. Then, we collect many generated patterns and apply the principal component analysis to $(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)$ and $(\boldsymbol{\sigma}_{xx}, \boldsymbol{\sigma}_{yy}, \boldsymbol{\sigma}_{xy})$ to extract the first principal components $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ of the same constant size T_{gen}. Finally, we denote the features of saliency primitive $D_k^{(c)}$ as $\hat{\Theta}_k^{(c)} = \left(\hat{\boldsymbol{\mu}}_k^{(c)}, \hat{\boldsymbol{\sigma}}_k^{(c)}, \hat{\boldsymbol{\phi}}_k^{(c)}\right)^{\mathrm{T}}$, where $\hat{\boldsymbol{\phi}}_k^{(c)}$ is obtained by normalizing $\boldsymbol{\phi}_k^{(c)}$.

4.3 Gaze-based Feature Extraction for Scene-level Correlations

Gaze-based feature extraction aims to describe the characteristics of scene-level correlations between scene structures and gaze dynamics with the object of "how specific types of gaze primitives can be characterized when looking at a certain type of saliency primitives". As considered in the previous chapter, gaze primi-

tives basically reflect spatiotemporal patterns of saliency primitives being focused on. Thus, we identify the types of gaze primitives based on those of saliency primitives of focus. To this end, we first classify the types of saliency primitives so that different types of gaze primitives can be observed according to the types of saliency primitives (see Section 4.3.1). Then, we extract different features for the types of gaze primitives in Section 4.3.2.

4.3.1 Classification of Saliency Primitives and Gaze Primitives

As introduced in [Jacob and Karn, 2003], many features extracted from gaze data are unique to eye movement types. We thus aim to classify gaze primitives so that each type of primitives corresponds to that of eye movements such as fixations and pursuits. To this end, we classify the types of saliency primitives based on the translation speeds of salient regions since the differences in the speeds provide different types of eye movements. Specifically, we utilize $\hat{\boldsymbol{\mu}}_k^{(c)}$ in $\hat{\Theta}_k^{(c)}$ to calculate the average motion speeds as $z_k^{(c)} = \| \max(\hat{\boldsymbol{\mu}}_k^{(c)}) - \min(\hat{\boldsymbol{\mu}}_k^{(c)}) \| / T_{\text{gen}}$. We collect many samples of translation speeds from various videos and conduct a standard k-mean clustering to derive two types of saliency primitives $\mathcal{W} = \{w_1, w_2\}$ where w_1 and w_2 correspond to static and dynamic salient regions, respectively.

Then, let us denote a gaze data in the k-th interval as $X_k = \left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{T_k}\right)$ where $T_k = i_{k2} - i_{k1} + 1$. In addition, all the saliency primitives provide sequence of positions; we describe the positions of regions at frame t as $\{\boldsymbol{\mu}_t^{(1)}, \ldots, \boldsymbol{\mu}_t^{(C_k)}\}$ where C_k is the number of the regions. As introduced in Section 4.2.2, the types of these primitives are identified as $W_k^{(c)} \in \mathcal{W}$. As well, we denote the type identified to the c-th region in frame t as $\omega_t^{(c)}$. For each frame, we first refer to saliency primitive being looked at, $\bar{c}_t$, based on the distances between the location of gaze points and that of regions: $\bar{c}_t = \arg\min_c \| \boldsymbol{p}_t - \boldsymbol{\mu}_t^{(c)} \|$. We then split gaze data where $\bar{c}_t \neq \bar{c}_{t+1}$ or $\omega_t^{(\bar{c}_t)} \neq \omega_{t+1}^{(\bar{c}_t)}$ to obtain a gaze primitive sequence, $G_k = (g_1, \ldots, g_{\kappa_{\text{all}}})$, where g_κ has properties of region ID $\bar{c}_\kappa \in \{1, \ldots, C\}$, primitive type of focus $\omega_\kappa \in \mathcal{W}$, and gaze data $X_\kappa = \left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{T_\kappa}\right)$ where T_κ is the size of the primitives.

Finally, we identify the types of gaze primitives based on those of saliency primitives being looked at, ω_κ. Specifically, g_κ obtains fixation labels if $\omega_\kappa = w_1$ and otherwise gets pursuit labels. In addition, we refer to state transitions from g_κ to $g_{\kappa+1}$ as saccades if $\bar{c}_\kappa \neq \bar{c}_{\kappa+1}$.

4.3.2 Feature Extraction

This section proposes a feature extraction method to obtain $\boldsymbol{\varphi}(X_k, \mathcal{D}_k)$ in Equation (4.2). X_k often contains multiple gaze primitives of different types as presented in Section 4.3.1, where the observable types depend on those of scene structures $hist(W_k)$. What we introduce here is an adaptive feature extraction method that extracts different features for the types of gaze primitives and aggregate them to derive $\boldsymbol{\varphi}(X_k, \mathcal{D}_k)$ based on scene structure types.

First, fixations contain internal gaze shifts to scan objects. We suppose that such shifts occur more actively when observers are in a higher level of attentiveness, and thus we introduce the size and the frequency of the shifts as features. Here, the gaze data of the κ-th segment is denoted as $X_\kappa = \left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{T_\kappa}\right)$. If the segment has a label of fixation, we denote a set of shifts as $\dot{X}_\kappa = \left\{\dot{\boldsymbol{p}}_1 \ldots, \dot{\boldsymbol{p}}_{T_\kappa - 1}\right\}$ where $\dot{\boldsymbol{p}}_t = \boldsymbol{p}_{t+1} - \boldsymbol{p}_t$. We then extract parts of the shifts from $\dot{X}_\kappa$ that are larger than threshold π_v, $\dot{X}'_\kappa = \{\dot{\boldsymbol{p}}_t \mid \|\dot{\boldsymbol{p}}_t\| > \pi_\mathrm{v}\}$ to calculate the following two features.

$$
\begin{aligned}
e_\mathrm{size} &= \frac{1}{|\dot{X}'_\kappa|} \sum_{\boldsymbol{p} \in \dot{X}'_\kappa} \|\boldsymbol{p}\|, \\
e_\mathrm{freq} &= \frac{|\dot{X}'_\kappa|}{T_\kappa},
\end{aligned}
$$

where $|\dot{X}'_\kappa|$ is the cardinality of $\dot{X}'_\kappa$.

As for features of pursuits, we extract the synchronization of speeds between gaze shifts and the motions of salient regions. When humans track a moving object, they tend to synchronize the pursuit acceleration to the expected changes of target motions and maintain the velocity at a constant level as long as the target velocity is not expected to change [Becker and Fuchs, 1985]. Let us assume a sequence of locations of the salient regions being looked at during X_κ as $\Theta_\kappa = \left(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{T_\kappa}\right)$, where the segment is given a label of pursuits. In addition, we denote a set of target motion speeds as $\dot{\Theta}_\kappa = \left\{\dot{\boldsymbol{\mu}}_1 \ldots, \dot{\boldsymbol{\mu}}_{T_\kappa - 1}\right\}$ where $\dot{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t$. Then, we calculate the synchronization features as follows:

$$
\begin{aligned}
e_\mathrm{sync} &= \frac{1}{T_\kappa - 1} \sum_{t=1}^{T_\kappa - 1} \frac{\|\dot{\boldsymbol{p}}_t\| \cos o_t}{\|\dot{\boldsymbol{\mu}}_t\|}, \\
e_\mathrm{res} &= \frac{1}{T_\kappa - 1} \sum_{t=1}^{T_\kappa - 1} \|\dot{\boldsymbol{p}}_t\| \sin o_t,
\end{aligned}
$$

where $\cos o_t = \frac{\dot{\boldsymbol{p}}_t \cdot \dot{\boldsymbol{\mu}}_t}{\|\dot{\boldsymbol{p}}_t\| \|\dot{\boldsymbol{\mu}}_t\|}$. Note that e_{res} is a feature to describe a residual element that implicitly indicates scanning behavior during pursuits.

In addition to the features presented above, we introduce features for saccades. Recall that the gaze data in interval I_k is split into a sequence of κ_{all} gaze primitives, where transitions between primitives are given by the changes of saliency primitives or their types being looked at (static to dynamic or the opposite). We here denote $\kappa'_{\text{all}} \leq \kappa_{\text{all}}$ as the number of transitions given by the changes of primitives. Since it is affected by the number of primitives in a scene structure as well as the length of the interval, we introduce the following two features.

$$\begin{aligned} e_{\text{sacn}} &= \kappa'_{\text{all}} / C_k, \\ e_{\text{sacl}} &= \kappa'_{\text{all}} / T_k. \end{aligned}$$

Finally, we aggregate the features introduced so far based on the types of scene structures. That is, feature $\boldsymbol{\varphi}(X_k, \mathcal{D}_k)$ includes $\bar{e}_{\text{size}}, \bar{e}_{\text{freq}}$ and $\bar{e}_{\text{sync}}, \bar{e}_{\text{res}}$ if the scene structures contain w_1 and w_2, respectively, where $\bar{e}$ indicates an average of the feature values extracted from all the gaze primitives being concerned. In addition, the feature involves $e_{\text{sacn}}, e_{\text{sacl}}$ when the scene contains multiple primitives.

4.4 Saliency-based Feature Extraction for Scene-level Correlations

The saliency-based feature extraction aims to describe "which types of saliency primitives originally tend to be looked at in a certain type of scene structures". In other words, we investigate what types of visual events tend to be looked at in the light of saliency. Although the procedure in Section 4.2.2 introduces a simplified representation of saliency primitives, it still indicates various types of visual events such as translations, resizes and the variations of saliency. Since it is difficult to introduce prior knowledge on which types of visual events frequently appear in videos and furthermore how much they tend to attract eyes, we introduce the classification of primitives that preserves all the properties as far as possible in Section 4.4.1. Then, we define the saliency-based feature of scene-level correlations in Section 4.4.2.

4.4.1 Classification of Saliency Primitive Types

Assume that N' saliency primitives $\mathcal{D}_{\text{all}} = \{D_1 \dots, D_{N'}\}$ are obtained from a set of videos, where D_n is characterized by $3 \times T_{\text{gen}}$ matrix feature $\hat{\Theta}_n = \left(\hat{\boldsymbol{\mu}}_n, \hat{\sigma}_n, \hat{\boldsymbol{\phi}}_n\right)^{\mathrm{T}}$. We cluster them into N types, $\mathcal{W} = \{w_1, \dots, w_N\}$, while fully utilizing all the properties in the patterns. In the clustering, we define a dissimilarity metric between primitives based on the correlation between two generated patterns. Specifically, we define the dissimilarity between D_n and $D_{n'}$ as follows:

$$Z(D_n, D_{n'}) = 1 - \frac{1}{3}\left(\text{ZNCC}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\mu}}_{n'}) + \text{ZNCC}(\hat{\sigma}_n, \hat{\sigma}_{n'}) + \text{ZNCC}(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\phi}}_{n'})\right),$$

where ZNCC describes a normalized cross correlation between patterns. With this similarity, we conduct a hierarchical clustering of saliency primitives into the predefined number of types, N (we will discuss a measure to determine N in the experiment section). Specifically, we adopt the complete linkage algorithm and give the maximum value of dissimilarities for a given pair of primitives $D_{u_1} \in U_1, D_{u_2} \in U_2$, $Z(D_{u_1}, D_{u_2})$, as the dissimilarity between two sets $U_1, U_2 \subset \mathcal{D}_{\text{all}}$.

As a result of the clustering presented above, we can visualize representative primitives for each type by identifying a single saliency primitive from spatiotemporal patterns of the same type. Figure 4.3 shows an example of representative primitives when $N = 5$. These representative primitives describe various visual events defined by the combinations of translations, resizes and variations of saliency. In addition, Figure 4.4 depicts selected identification results of saliency primitive types corresponding to Figure 2.6, where the colors correspond to the types in Figure 4.3. Although representative primitives in Figure 4.3 do not always describe the original primitives in Figure 2.6 accurately when N is small, we can still classify scene structures into several types based on the combination of saliency primitive types in a data-driven manner.

4.4.2 Feature Extraction

This section introduces the description of feature $\boldsymbol{\varphi}(\boldsymbol{p}_t, \mathcal{D}_k)$ in Equation (4.3). Specifically, we utilize spatial locational relationships between saliency primitives and gaze points to learn the types of saliency primitives that tend to be looked at in a soft-assignment fashion.

In interval I_k, we have scene structure $\mathcal{D}_k = \{D_k^{(1)}, \dots, D_k^{(C_k)}\}$ where $D_k^{(c)}$ is identified type $W_k^{(c)} \in \mathcal{W} = \{w_1, \dots, w_N\}$. In addition, $D_k^{(c)}$ is located at $\boldsymbol{\mu}_t^{(c)}$ in

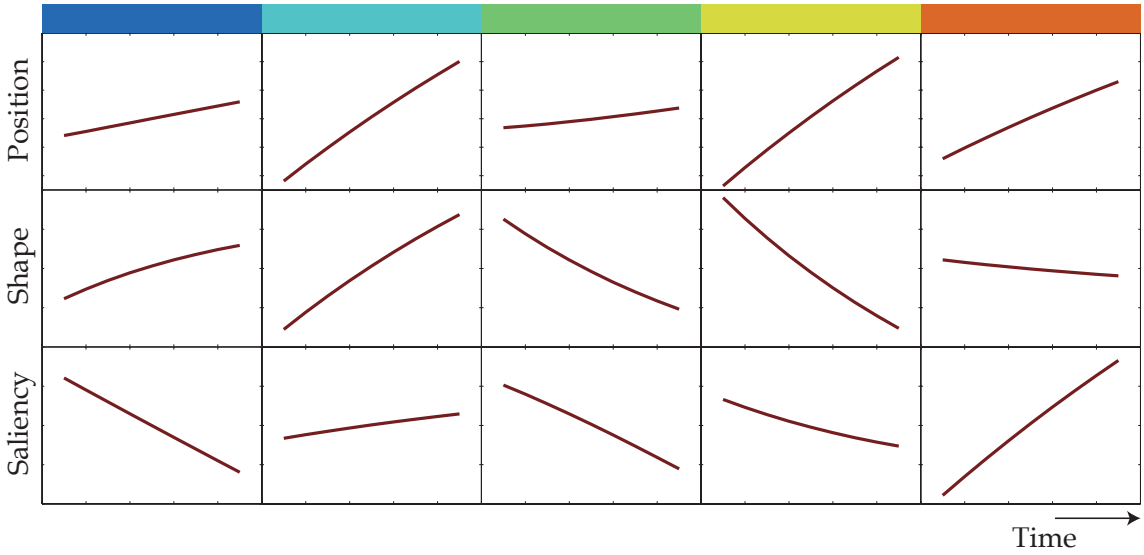

Figure 4.3: Example of representative primitives for each type ($N = 5$).

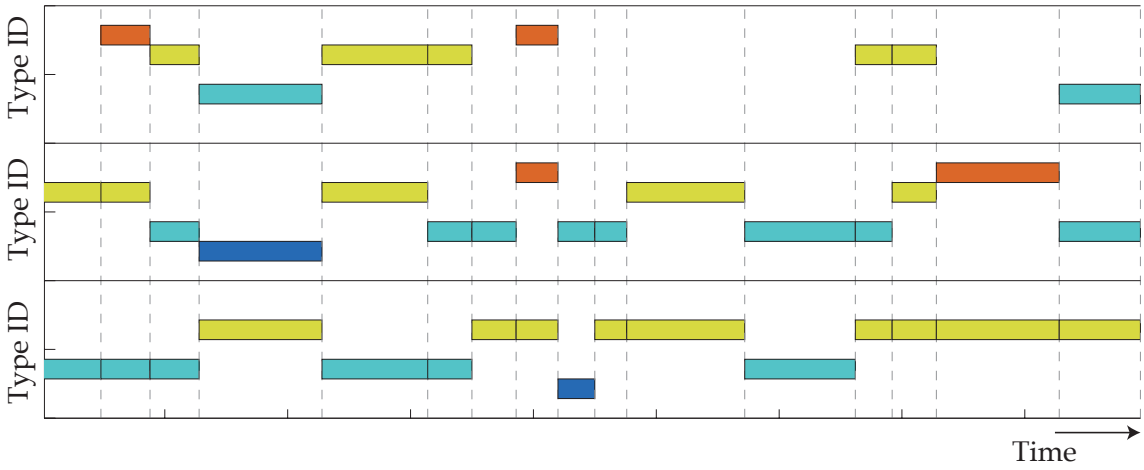

Figure 4.4: Example of identified saliency primitive types corresponding to Figure 2.6. The colors and vertical positions of each rectangle describe the ID of identified types, where the colors correspond to Figure 4.3. The combination of types in each interval (split by dotted lines) defines the types of scene structures.

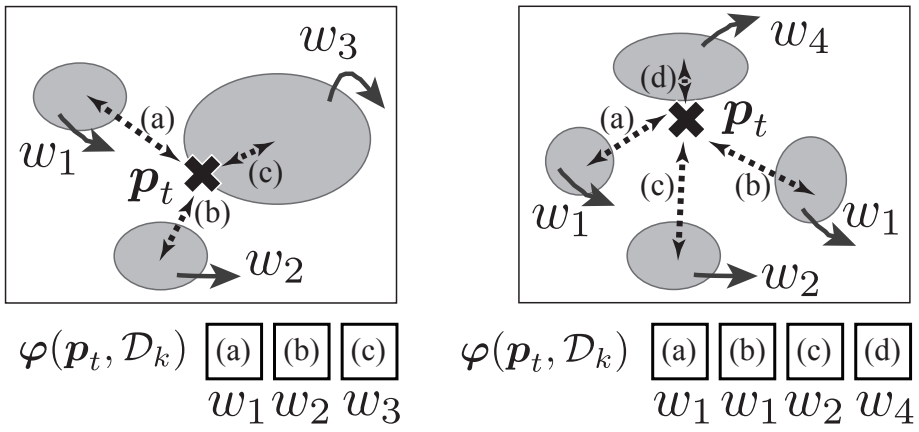

Figure 4.5: Extracting spatial locational relationships between saliency primitives and gaze points for features.

frame t. First, we align $\mathcal{D}_k$ so that $D_k^{(c)}$ is lined up based on an ascending order of the types. In the cases where duplicate types given to a subset of $\mathcal{D}_k$, we line up the subset in a raster manner. We then describe the spatial locational relationships between saliency primitives and gaze points (see Figure 4.5). That is, we give a set of distances between the locations of saliency primitives and those of gaze as feature $\boldsymbol{\varphi}(\boldsymbol{p}_t, \mathcal{D}_k)$. More specifically, $\boldsymbol{\varphi}(\boldsymbol{p}_t, \mathcal{D}_k)$ is given as follows:

$$\boldsymbol{\varphi}(\boldsymbol{p}_t, \mathcal{D}_k) = \left(\|\boldsymbol{p}_t - \boldsymbol{\mu}_t^{(1)}\|, \ldots, \|\boldsymbol{p}_t - \boldsymbol{\mu}_t^{(C_k)}\| \right). \tag{4.4}$$

Equation (4.4) can provide a scene-adaptive but universal feature description since the dimension of feature vectors as well as the meaning of each feature dimension differ for the types of scene structures.

4.5 Experiments

In this chapter, we recorded gaze data of 10 subjects during watching TV commercial films in several conditions of attentiveness. Note that the videos were the same ones utilized in Section 2.3.5, which were intentionally designed to contain several distinct objects so that they were suitable for evaluating the proposed methods with the OSDM. In addition, since these videos contained frequent scene changes, subjects' gaze was basically expected to concentrate on salient regions that attracted their exogenous attention, although some endogenous actions like examinations and switches of attentional targets were likely to occur along with the experiments conducted in Chapter 3.

4.5.1 Experimental Setups

Experimental conditions

The objective of this chapter is to estimate attentive states that classifies if subjects concentrate on displayed videos or not. More specifically, we define attentive states as a state that specifies whether observers concentrate on a certain task (i.e., video-viewing task), which can be quantified into several levels. In this context, [Kahneman, 1973] has proposed the attention theory that likens attention to a limited resource which is allocated to tasks. Following this theory, the level of attentiveness can be regarded as the amount of attention resource allocating to the tasks. We therefore gave subjects the following instructions so that they were able to freely watch videos as far as possible in high/low level of attentiveness.

Task 1 (high level of attentiveness) : Please watch a video and answer the questionnaire to evaluate how much you liked the video on a seven-point scale.

Task 2 (low level of attentiveness) : Please watch a video while doing the following calculation task at the same time; please keep on subtracting 7 from 1000 and report answers (1000, 993, ...) to the experimenter. Please remember that you may have to repeat the session again in the cases where you reported too many wrong answers.

Tasks 1 and 2 corresponded to high and low attentive states, respectively. Above all, Task 2 made subjects conduct a secondary task (i.e., the calculation) to decrease the attention resource to the video-viewing task, where the caution about wrong answers was aimed at making the subjects focus on the secondary task.

Design of experiments

10 subjects individually sat in front of a screen[(i)], and a gaze tracking system[(ii)] was installed below the screen. The gaze-tracking accuracy was, on average, around 0.7°. The distance between the subject and the screen was around 1000 mm so that gaze dynamics were able to be observed during experiments.

12 TV commercials were split into two groups, VA (six out of all the videos) and VB (the other six videos). In addition, we also split 10 subjects into two groups, SA (five out of all the subjects) and SB (the other five subjects). Then, each

[(i)]MITSUBISHI Diamondcrysta RDT262WH, 25.5 inch, W550 mm/H344 mm.
[(ii)]Tobii X60 Eye Tracker. An approximate allowed range of head motion is 400×220×300 mm.

Table 4.1: Experimental procedures

Subject group SA		Subject group SB	
1st trial	Video group VA — Task 1	1st trial	Video group VB — Task 1
2nd trial	Video group VB — Task 2	2nd trial	Video group VA — Task 2
Short break and the recalibration of a gaze tracking system			
3rd trial	Video group VB — Task 1	3rd trial	Video group VA — Task 1
4th trial	Video group VA — Task 2	4th trial	Video group VB — Task 2

subject watched all the videos twice by following the procedures in Table 4.1. In each trial, the order of video playback was randomized for each of the subjects.

Preprocessing and preliminary evaluations

Gaze data was obtained at 30 fps, the same as the frame rate of videos being used. As a preprocessing, we applied a median filter with 0.5 sec window to the data to suppress spontaneous noises and to interpolate short defects by eye blinks. We also exclude the remaining defects in the data caused by eyelid closures from analyses, which constituted 23.6 % of the total data. The average and standard deviation of the scores obtained by the questionnaire in Task 1 were 4.74 and 1.18, respectively. With regard to Task 2, subjects reported the answers of calculations at least four times in each session.

4.5.2 Evaluation Scheme

We implemented and evaluated the gaze-based feature extraction in Section 4.3 ($\mathbf{M}_{\mathrm{G}}$) and the saliency-based extraction ($\mathbf{M}_{\mathrm{S}}$) in Section 4.4. The parameters for the OSDM were configured as the same as those in Section 2.3.5. As for parameter π_{v} to extract internal gaze shifts in Equation (4.3.2), we empirically set $\pi_{\mathrm{v}} = 0.1$ in 80×60 pixel frames (approximately 1°/sec). In the preliminary experiments, we confirmed that π_{v} did not affect final estimation accuracies significantly after smoothing gaze data by the preprocessing presented above. T_{gen} to classify the types of saliency primitives was empirically set to $T_{\mathrm{gen}} = 0.5$ sec (15 frames) so as not to generate similar patterns regardless of the underlying parameters of AR models. In addition to the two proposed methods, we implemented the baseline method that followed the formulation in Equation (4.1) and did not particularly use scene structure information. Specifically, we aggregated all the features $\bar{e}_{\mathrm{size}}, \bar{e}_{\mathrm{freq}}, \bar{e}_{\mathrm{sync}}, \bar{e}_{\mathrm{res}}, e_{\mathrm{sacn}}, e_{\mathrm{sacl}}$ for $\boldsymbol{\varphi}(X_k)$ regardless of the saliency primitive types

Table 4.2: Estimation results

Method	Baseline	M_G	M_S ($N = 5$)	M_S ($N = 6$)	M_S ($N = 7$)
Accuracy [%]	66.4	70.2	78.7	78.6	81.8
# types	1	13	46	50	64
Coverage [%]	100	100	52.2	48.0	45.3

of focus and learned them without distinction of scene structure types.

We evaluated our methods based on the leave-one-out cross validation scheme. Specifically, we divided 240 sequences consisting of 10 subjects $\times$ 12 videos $\times$ 2 conditions into 239 training and 1 test sequences. Naive Bayes classifiers were adopted to train Equation (4.1), (4.2) or (4.3). In the training, obtained distributions of feature values were smoothed via the kernel density estimation, where the parameter of kernel was tuned by the cross validation in training sequences. Although M_S obtains estimation results per frame, we vote frame-wise results within each interval to derive estimation results per interval for the sake of fairness to M_G and the baseline method. Finally, an estimation accuracy was obtained as the ratio of intervals that were given correct attentive states.

Since we conduct the estimation based on a supervised learning framework, the proposed methods can be only applicable to trained types of scene structures. That is, we cannot conduct the estimation if features were not learned for the types of scene structures observed in a test video. Although we used all the videos for the training, we need to estimate attentive states for unseen videos in a practical situation. Thus, we counted the number of scene structure types in the 12 videos (**# types** in the next section) and evaluated the ratio of scene structure types that appeared in more than one videos (**Coverage**) as a measure to evaluate a generalization capability on the unseen videos. In addition, we tested several numbers of saliency primitive types in the M_S, N in Section 4.4.1, and compared them in terms of the accuracy and coverage.

4.5.3 Results and Discussions

Comparison between gaze-based and saliency-based feature extraction

Table 4.2 described the scores of all the methods. These results demonstrate the effectiveness of utilizing scene-level correlations in terms of predicting attentive states. Among M_S with different numbers of primitive types, N, we can obtain

high accuracies by M_S if we set large N. On the other hand, the number of scene structure types increases and the coverage decreases as N becomes larger. To determine appropriate N based on these different measures, we introduce a harmonic mean[(iii)] of the following form:

$$F = \frac{2 \cdot \text{Accuracy} \cdot \text{Coverage}}{\text{Accuracy} + \text{Coverage}}. \tag{4.5}$$

Among M_S with different N, harmonic mean F showed the best when $N = 5$ ($F = 0.628$) and monotonically decreased as N becomes large. On the other hand, M_G obtained $F = 0.825$, which was significantly higher than any other M_S. In conclusion, the saliency-based feature extraction works better if we can assume all the videos are given and trained preliminary, while the gaze-based extraction has the advantage of being applicable to unseen videos.

Figure 4.6 depicts selected examples of estimation results. In the 1st and 2nd columns, color points show subsequences of gaze points (gaze points at ± 3 frames) for all the subjects, where red and yellow show high and low attentive states, respectively. The 3rd column contains fitting results of the OSDM. When subjects looked at different regions for the levels of attentiveness, the saliency-based features work effectively as shown in Examples (A) and (B). The 4th and 5th columns of these examples depict selected properties of saliency primitives where the color of lines shows the types of the primitives described in Figure 4.3 ($N = 5$). In Example (A), gaze points under the high level of attentiveness (red) concentrated on the 3rd and 4th saliency primitives. These primitives correspond to the appearance event of an object with a large translation, while the other primitives being looked at under the low level of attentiveness describe smaller translation events. Example (B) has two saliency primitives in its scene structure, and the distributions of gaze points differ for the levels of attentiveness. The 2nd region, which tended to be looked at more frequently when subjects were in the high level of attentiveness, corresponds to a text caption with visual events of losing saliency due to an appearance event of a new object from the top of frame. Although the semantic meaning of region (i.e., text caption) is invisible in the proposed method, we can capture the tendency of gaze behavior from the viewpoint that what types of saliency primitives are attracting eyes.

[(iii)] A harmonic mean is originally aimed at measuring an average performance of different measures. For example, the F-measure in the field of information retrieval is the harmonic mean of precision and recall scores.

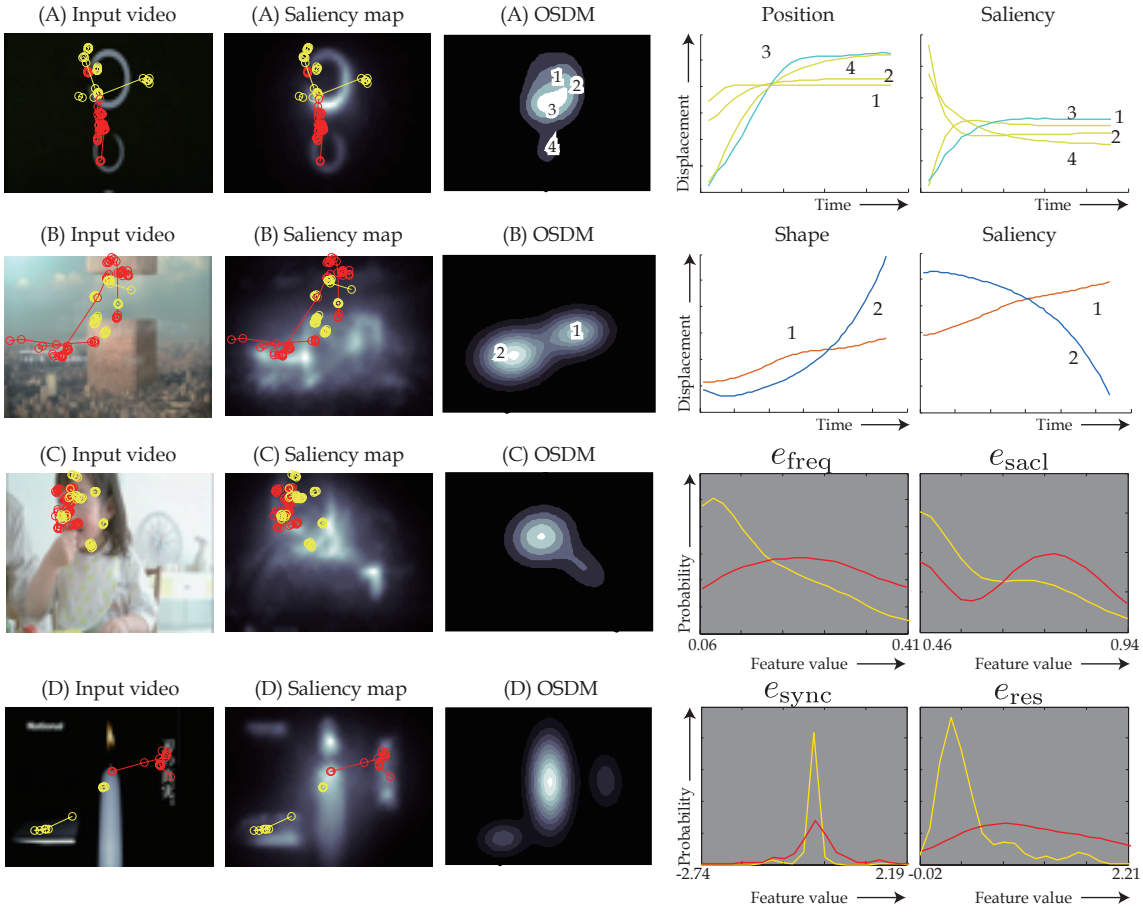

Figure 4.6: Estimation results. 1st column: input videos, 2nd column: saliency maps, and 3rd column: fitting results of the OSDM. The red and yellow points indicate subsequences of gaze points (gaze points at ± 3 frames) for all the subjects under high and low attentive states, respectively. In Examples (A) and (B), the 4th and 5th columns depict selective properties of saliency primitives shown in the titles, where the numbers from the 3rd to 5th columns indicate the ID of saliency primitives. In addition, the color of lines are the ID of the primitive types described in Figure 4.3. In Examples (C) and (D), the 4th and 5th columns describe the estimated probability distributions for the selected gaze features shown in the titles, where the color of lines correspond to the points of gaze in the 1st and 2nd columns. The images used in this figure were provided by courtesy of Panasonic Corporation.

Examples (C) and (D) show the estimated probability distributions of several gaze-based features that contributed to the estimation. Since gaze points in Example (C) concentrated on a face regardless of attentive states, it is difficult to introduce the saliency-based method for this situation. However, there were differences in the frequency of gaze shifts when fixating the face and that of saccades as shown in the 4th and 5th columns of the example. Specifically, subjects tended to provide gaze shifts and saccades more frequently when they were highly attentive. Example (D) describes another tendency of gaze behavior when pursuing objects with translation events. As shown in the 4th and 5th columns, subjects tended to pursue moving targets with a more constant ratio of speeds and directions in when they were in the low level of attentiveness. Alternatively, subjects tended to examine objects with translation events more actively when they were highly attentive.

Integration of gaze-based and saliency-based feature extraction

The discussions in the preceding section indicate that we can adopt an appropriate feature extraction from the saliency-based and gaze-based ones on the following two criteria:

(1) Types of problems to solve Which of the estimation accuracy and coverage (generalization capability on unseen videos) we should consider.

(2) Characteristics of scene structures. The number of regions, for example. If there is only a single object attracting our attention, such as Example (C) in Figure 4.6, we cannot use the saliency-based feature extraction.

On the other hand, we can introduce a more balanced estimation by integrating the two feature extraction methods based on the accuracy and coverage obtained in the experimental results. Here, we propose simple late-integration techniques for each video. Let us denote a sequence of estimation results when the i-th subject is looking at a video consisting of K intervals in attentive state $a \in \{A_1, A_2\}$ as $\boldsymbol{r}_{ia}^{(\mathrm{g})} = (r_{ia1}^{(\mathrm{g})}, \ldots, r_{iaK}^{(\mathrm{g})}) \in \{0,1\}^K$ for $\mathrm{M_G}$ and $\boldsymbol{r}_{ia}^{(\mathrm{s})} = (r_{ia1}^{(\mathrm{s})}, \ldots, r_{iaK}^{(\mathrm{s})}) \in \{0,1\}^K$ for $\mathrm{M_S}$, where $i \in N_{\mathrm{subject}} = \{1 \ldots, 10\}$ is the ID of subjects, $r_{iak}^{(\mathrm{s})} = 1$ indicates correct estimates and $r_{iak}^{(\mathrm{s})} = 0$ otherwise. In addition, we prepare a sequence of coverage indicators $\boldsymbol{q}_{ia}^{(\mathrm{s})} = (q_{ia1}^{(\mathrm{s})}, \ldots, q_{iaK}^{(\mathrm{s})}) \in \{0,1\}^K$ in the same form, where $q_{iak}^{(\mathrm{s})} = 1$ if the type of scene structures in this interval appears only in this video. On that basis, we introduce the following two integration techniques.

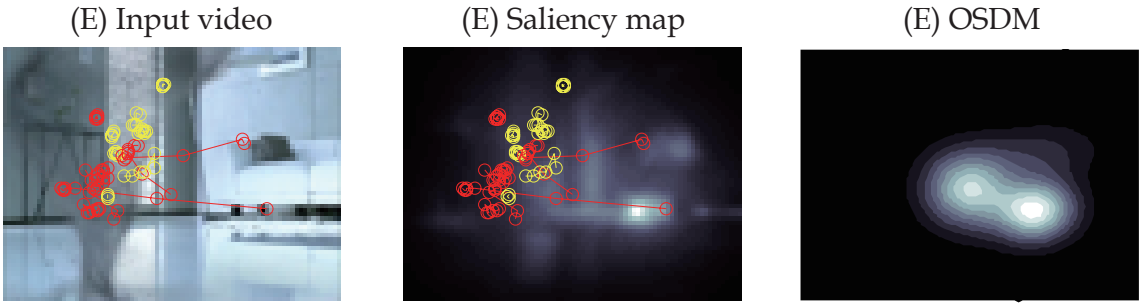

Figure 4.7: Failure case. The images used in this figure were provided by courtesy of Panasonic Corporation.

Accuracy-based integration. Given the data above, we split them into 1 test and the remaining training data (let's say i_{test} for the ID of a test subject) and vote the estimation results of the training data with respect to each feature extraction: $r_k^{(m)} = \sum_a \sum_{i \in N_{\text{subject}} \setminus i_{\text{subject}}} r_{iak}^{(m)}$, where $m \in \{\text{g}, \text{s}\}$. Based on the votes, we can select feature extraction methods to use in the k-th interval: $m_k = \arg\max_m r_k^{(\text{m})} \in \{\text{g}, \text{s}\}$. Finally, we integrate $\boldsymbol{r}_{i_{\text{test}}a}^{(\text{g})}$ and $\boldsymbol{r}_{i_{\text{test}}a}^{(\text{s})}$ by referring to the selected extraction methods with respect to each interval based on $(m_1, \ldots, m_K)$.

We applied the above method to the obtained results where M_S was configured with $N = 5$. Since M_S was selected in most of the sequences (the ratio of intervals choosing M_S was 70.2%, while the ratio for M_G was 22.2% and the remaining 7.6% of all the intervals had the same scores between the feature types), the estimation accuracy was slightly improved to 80.1% by the integration.

Coverage-based integration. In this integration, we refer to $\boldsymbol{q}_{ia}^{(\text{s})} = (q_{ia1}^{(\text{s})}, \ldots, q_{iaK}^{(\text{s})})$ to utilize result $r_{iak}^{(\text{g})}$ instead of $r_{iak}^{(\text{s})}$ in the k-th interval if $q_{iak}^{(\text{s})} = 1$. Namely, this integration method relies on M_S if the training results can be inherited from other videos and otherwise adopts M_G. The estimation accuracy by the coverage-based integration method with M_S ($N = 5$) was 73.4%, which demonstrated fair amount of improvement from M_G in Table 4.6 while maintaining the coverage at 100%. The harmonic mean defined in Equation (4.5) was also improved to $F = 0.847$.

Limitations of the developed framework

The experimental results also indicate several limitations of the proposed framework. Although the TV commercial films that we used in the experiments have

the potential to provide more exogenous motions than endogenous actions as mentioned in the beginning of Section 4.5, subjects were sometimes attracted to non-salient regions, which are invisible and thus a limitation in our framework. For example, Figure 4.7 describes a typical failure case of the proposed method. In Example (E), most of the gaze points are directed to the object (fugitive dust) which is not salient but semantically conspicuous.

In addition to the preceding limitation, we did not take any acoustic information into account during the analyses. For example, commercial films often contain narration speeches that explain products and logos. For such cases, objects associated with the acoustic information should be given a particularly higher saliency. These limitations posed here are basically related with the modeling of saliency dynamics, and we will discuss them in Chapter 6.

Chapter 5

Gaze Point Prediction from Spatiotemporal Correlations

5.1 Introduction

In the previous chapters, we focused on event-level spatiotemporal gaps and scene-level correlations separately based on the proposed framework. While we analyzed the spatiotemporal gaps provided by a single type of gaze primitives in Chapter 3, the degree of gaps can vary depending on the types of gaze primitives; in Figure 2.10, a large gap occurred particularly when gaze shifted larger. Moreover, the gaps are also influenced by the types of visual events and furthermore, the time-varying scene structures like Chapter 4. For example, sudden motions of objects among many static objects can provide a large reaction delay. Consequently, the event-level spatiotemporal gaps can be influenced by the scene-level correlations consisting of scene structures and gaze dynamics. The aim of this chapter is to describe the overall spatiotemporal correlations of the aforementioned characteristics based on the proposed framework.

As a practical situation where gaze behavior exhibits the spatiotemporal correlations, we assume a free-viewing of a more variety of videos than previous chapters, including unedited natural ones such as surveillance videos. Since those videos do not always contain distinct objects that can be easily followed by observers nor frequent scene changes, eyes can be sometimes directed to irrelevant locations. Figure 5.1 depicts an example of the above situation. Although the video displays a bus that can be a salient region, several gaze points (depicted as red points) could not follow it and provided a gap since the bus contained a fast translation event from the left to the right of a frame. Moreover, several points

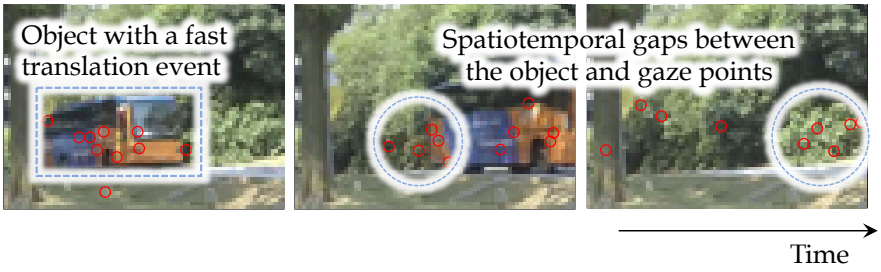

Figure 5.1: Example of spatiotemporal correlations when watching videos. Parts of the photos in this figure are contained in the dataset provided by [Li et al., 2004]. Red points indicate ground truths of gaze points, where each point corresponds to one individual subject in [Riche et al., 2012].

remained the right of the frame even if the bus has disappeared. One of the possible interpretations for such situations is a temporal delay in reactions. Observers might try to orient their overt attention to the bus but failed and unconsciously look at background regions. Alternatively, they might look at the background regions consciously because they tried to get some information from those regions. Obviously, this example results from the spatiotemporal correlations. Although we cannot determine which of the above two interpretations can actually explain their gaze behavior, we can see an event-level spatiotemporal gap reflecting a scene-level correlation consisting of specific saliency and gaze dynamics (i.e., the fast translation event and reaction pursuit in this example).

Towards the description of overall spatiotemporal correlations, we first introduce a model to describe the relationships between spatiotemporal gaps and scene structures that influence the gaps, which we refer to as *gap structures*. Specifically, we leverage saliency primitives of the patch-based saliency dynamics model (PSDM) to describe both gaps and scene structures jointly (see Figure 5.2). Then, we statistically learn the modeled gap structures around the points of gaze for each type of gaze primitives so that we can involve their scene-level correlations with gaze dynamics. Intuitively, we learn the gap structures for fixations, pursuits and saccades individually. Finally, the learned relationships between gap structures and gaze primitive types describe overall spatiotemporal correlations consisting of event-level spatiotemporal gaps and scene-level correlations.

We leverage the proposed description for the task of gaze point prediction from videos. Gaze-point prediction techniques are applicable to not only proficiency estimation [Eivazi et al., 2012] and detection of developmental disor-

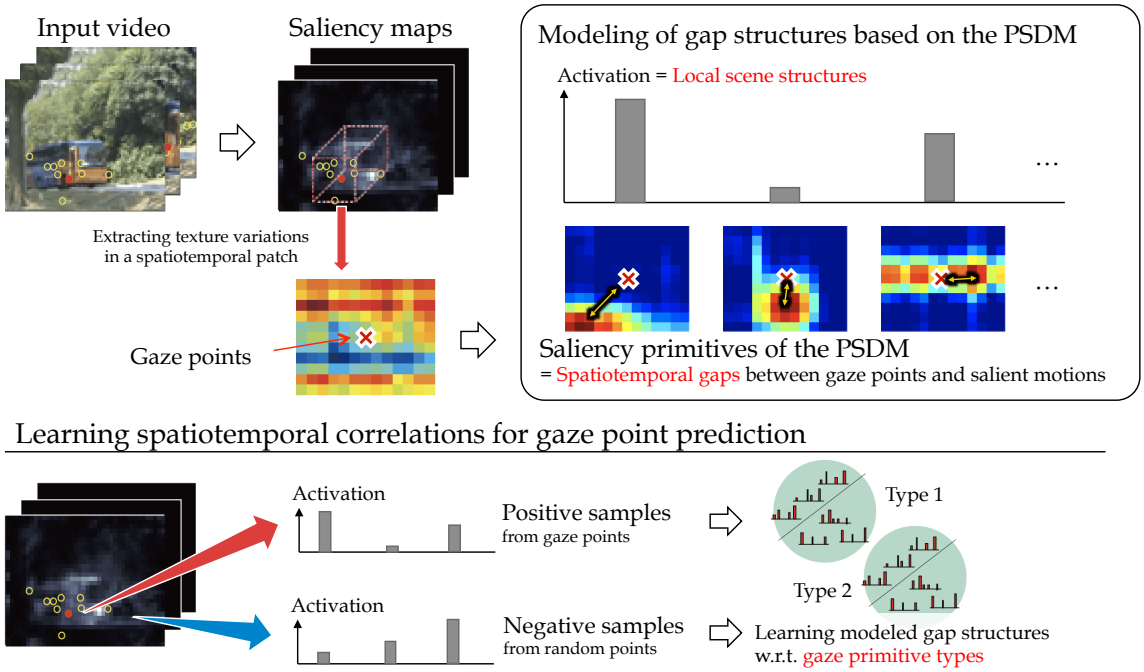

Figure 5.2: Describing spatiotemporal correlations based on the gap structure model. Parts of the photos in this figure are contained in the dataset provided by [Li et al., 2004].

ders [Tseng et al., 2013] from eye movements but gaze-based content designs and recommender systems [Simonin et al., 2005, Yoshitaka et al., 2007]. These applications all need to know or to predict where humans actually look rather than to extract informative regions from videos. While we follow traditional learning-based saliency maps (LBSM) that just predict if a certain point tends to be looked at by learning discriminative or regression models (e.g., [Peters and Itti, 2007, Judd et al., 2009, Kienzle et al., 2009, Borji, 2012]), the proposed method is novel in terms of (1) predicting gaze while considering its spatiotemporal gaps such as reaction delays and anticipation, and (2) predicting gaze while considering the type of gaze primitives. Note that the proposed method is also different from spatiotemporal saliency reviewed in 2.1.2 that aims to predict attention from motions since our method predicts how much gaps can be provided due to the motions.

In the following sections, we first introduce a traditional formulation of gaze point prediction in Section 5.2.1. Then, we introduce gap structure models and present how to involve the scene-level correlations with gaze primitive types in Section 5.2.2. In addition, we also propose another modeling technique of gap structures without utilizing the PSDM in Section 5.2.3.

5.2 Proposed Method

5.2.1 Gaze Point Prediction

The gaze point prediction is a task to predict where humans tend to look in each frame of videos. More specifically, we are to generate a prediction map where each pixel-value indicates the degree of gaze-point existence.

To address this task, we basically follow the traditional LBSM approach like [Peters and Itti, 2007, Judd et al., 2009, Kienzle et al., 2009, Borji, 2012]. The LBSM originally involves a supervised learning framework with a set of saliency-related features and gaze data as a training dataset. Let us denote a point of gaze in a dataset as $\boldsymbol{p} \in \mathbb{R}^2$ and features extracted from $\boldsymbol{p}$ as $\boldsymbol{\varphi}(\boldsymbol{p}) \in \mathbb{R}^{N_{\text{feat}}}$, where N_{feat} is the number of features. The LBSM aims to provide the degree of gaze-point existence at all the pixels as a continuous value, $B(\boldsymbol{p}) \in \mathbb{R}$. Namely, it predicts where observers tend to look in a map form for each video frame. We refer to the map as a *gaze-prediction map* to distinguish it from saliency maps. Since videos contain multiple frames, the final output is a sequence of gaze-prediction maps.

As for a model of $B(\boldsymbol{p})$, we introduce the following linear function:

$$B(\boldsymbol{p}) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{p}), \tag{5.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{N_{\text{feat}}}$ is parameters of the model. When we can prepare a distribution of gaze points (often called a *heat map*) for each image in a training dataset, that means we have a ground truth of $B(\boldsymbol{p})$ and can estimate parameter $\boldsymbol{\beta}$ directly via linear regression [Peters and Itti, 2007, Borji, 2012]. However, when dealing with videos, we cannot expect a dense heat map since the number of points in each frame is at most the number of subjects. For that case, we estimate $\boldsymbol{\beta}$ in a discriminative model [Kienzle et al., 2009, Judd et al., 2009, Borji, 2012]; $\boldsymbol{p}$ in the training dataset is given a label consisting of $\{1, -1\}$, where 1 is the positive label indicating the point tends to be looked at and -1 corresponds to a negative label for the little probability of being looked at. Then, $\boldsymbol{\beta}$ can be trained as parameters of discriminant function $B'(\boldsymbol{p})$ of the following form:

$$B'(\boldsymbol{p}) = \operatorname{sgn}(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{p}) + \beta_0), \tag{5.2}$$

where $\beta_0 \in \mathbb{R}$ is a bias term. Note that we can introduce non-linear classifiers instead of Equation (5.2) like [Kienzle et al., 2009]. Generally, the advantage of lin-

ear classifiers includes their computation speed and ease of interpretations while non-linear classifiers are slower but more powerful in classification.

Positive samples are often collected from the points of gaze in a training dataset. On the other hand, since salient regions that can attract our eyes are generally known to be sparse (e.g., see [Sun et al., 2012]), the negatives can be practically collected from random points.

5.2.2 Introducing Spatiotemporal Correlations

In this study, we introduce a model of gap structures that describe the relationships between event-level spatiotemporal gaps and scene structures. We exploit the modeled gap structures for feature description $\boldsymbol{\varphi}(\boldsymbol{p})$ in Equation (5.1) to predict gaze while considering time-varying scene structures in videos as well as spatiotemporal gaps that can appear when looking at specific visual events. In addition, by learning the discriminative model presented in Equation (5.2) with respect to each type of gaze primitives, we can also introduce the scene-level correlations between scene structures and gaze dynamics. We first discuss how to model the gap structures, and then extend the framework furthermore to take into account of the types of gaze primitives as well.

Modeling gap structures

We first introduce the assumption that the degree of gaze-point existence is particularly influenced by visual events around the points of gaze. Such an assumption can be often seen in traditional studies on saliency maps such as [Itti et al., 1998] that refer to local center-surround contrasts of visual stimuli. On that basis, we consider a local scene structure defined in a certain spatiotemporal patch (such as $\mathcal{N}(\boldsymbol{p}, t) := \Omega_{(\delta_x,\delta_y)} \times \mathcal{T}_{\delta_t}$ in Equation (2.3)) when introducing gap structures.

Specifically, the gap structures indicate what types of salient motions can be observed in a local scene structure around gaze points and how much spatiotemporal gaps appear against those motions. As a bottom-up approach to the modeling of gap structures, we utilize saliency primitives of the PSDMs presented in Section 2.4. In the PSDM, saliency primitives in codebook $\mathcal{D} = \{D_1, \ldots, D_N\}$ describe localized texture variations of saliency in a spatiotemporal patch. In other words, they indicate motion patterns and relative positions of salient regions. Then, given gaze point $\boldsymbol{p}$ as a center point of the patch, activation vector $\boldsymbol{w}(\boldsymbol{p}) = (w_1, \ldots, w_N)^{\mathrm{T}}$ describes local scene structures around gaze points

while each primitive contains spatiotemporal distances between the gaze points and salient motions (in what follows, we omit subscription t from the original definition of $\boldsymbol{w}(\boldsymbol{p}, t)$ and $\mathcal{N}(\boldsymbol{p}, t)$ without loss of generality). As for gaze point prediction, we utilize the activation vector $\boldsymbol{w}(\boldsymbol{p})$ as feature vector $\boldsymbol{\varphi}(\boldsymbol{p})$. Namely, the estimation of $\boldsymbol{\beta}$ can be regarded as a problem of finding the specific types of saliency primitives from codebook $\mathcal{D} = \{D_1, \ldots, D_N\}$ which have different tendencies in their appearances between the points of gaze and random points.

Here, the number of salient regions and their motion patterns can be different for patch sizes as shown in Figure 2.10. Although they can all attract our attention, we cannot know which ones actually influence gaze dynamics. We thus jointly consider multi-scale local scene structures from different neighborhoods of $\xi_{\max}$ scales, $\mathcal{N}_\xi(\boldsymbol{p})$ $(\xi = 1, \ldots, \xi_{\max})$. Specifically, saliency dynamics patterns in $\mathcal{N}_\xi(\boldsymbol{p})$ are first described with $L_{\mathcal{N}_\xi(\boldsymbol{p})}$ and its vectorized version with $\boldsymbol{l}_{\mathcal{N}_\xi(\boldsymbol{p})}$. We individually learn a codebook of primitives with respect to each scale, $\mathcal{D}_1, \ldots, \mathcal{D}_{\xi_{\max}}$, after resizing each $\mathcal{N}_\xi(\boldsymbol{p})$ into the same patch size. Then, activation vectors for scale ξ is described with $\boldsymbol{w}_\xi(\boldsymbol{p}) \in \mathbb{R}_+^{N_\xi}$ where N_ξ is the codebook size for scale ξ. Finally, we simply concatenate $\boldsymbol{w}_\xi(\boldsymbol{p})$ to derive feature vector $\boldsymbol{\varphi}(\boldsymbol{p})$, such as $\boldsymbol{\varphi}(\boldsymbol{p}) = (\boldsymbol{w}_1(\boldsymbol{p})^{\mathrm{T}}, \ldots, \boldsymbol{w}_{\xi_{\max}}(\boldsymbol{p})^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}_+^{N'}$ where $N' = \sum_\xi N_\xi$.

Incorporating the types of gaze primitives

To involve the scene-level correlations between scene structures and gaze dynamics, we statistically learn the modeled gap structures (which are embedded in $\boldsymbol{\varphi}(\boldsymbol{p})$ in gaze point prediction) with respect to each type of gaze primitives. Then, we calculate gaze prediction maps for all the types of gaze primitives, which individually indicate where humans tend to look with a certain gaze primitive type. Considering the existing findings that the gaps can vary depending on the types of eye movements, this approach is crucial since we can avoid learning a single model from gap structures with various tendencies.

In the proposed method of gaze point prediction, the obtained maps of each primitive type are finally integrated into single gaze-prediction maps. As a simple approach, we introduce the assumption that each type of gaze primitives can be observed with equal probability, independently and identically for spatial and temporal directions. Specifically, let us first denote the types of gaze primitives as $\mathcal{E} = \{e_1, \ldots, e_{N_{\mathrm{etype}}}\}$, where N_{etype} is the number of the types. By identifying gaze primitive types to each gaze point, $\boldsymbol{p}$ is given a label $g(\boldsymbol{p}) \in \mathcal{E}$ if $\boldsymbol{p}$ is a point of gaze and otherwise it is given a negative label. We then train Equation (5.2)

with respect to each type of gaze primitives from positive samples with label e_w and negative samples collected from random points to estimate parameters $\boldsymbol{\beta}_{e_1}, \ldots, \boldsymbol{\beta}_{e_{N_{\text{etype}}}}$. As a result, the degree of gaze-point existence with gaze primitive type e_w is evaluated as $B_{e_w}(\boldsymbol{p}) = \boldsymbol{\beta}_{e_w}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{p})$. Finally, we integrate model outputs over e_w to obtain the degree of gaze point existence:

$$B_E(\boldsymbol{p}) = \frac{1}{N_{\text{etype}}} \sum_{w=1}^{N_{\text{etype}}} B_{e_w}(\boldsymbol{p}). \tag{5.3}$$

By evaluating Equation (5.3) for all the pixels in the all of the frame of newly-observed videos, we finally obtain a sequence of gaze-prediction maps considering the types of gaze primitives.

Identification of gaze primitive types

In Chapter 4, we identified types of gaze primitives based on observed types of saliency primitives since we assumed overt attention was basically oriented to salient regions. However, this assumption is not always appropriate for the current situation since subjects can look at irrelevant locations unconsciously as mentioned in the beginning of this chapter. We therefore take a bottom-up approach to identify the types of gaze primitives. Specifically, we classify the types based on the motion speeds of gaze shifts in each gaze primitive.

Let us introduce gaze data $X = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T)$. First, X is applied a sliding temporal window to extract their subsequences of the fixed length τ_{fix}, $X_t = \left(\boldsymbol{p}_t, \ldots, \boldsymbol{p}_{t+\tau_{\text{fix}}-1}\right)$. We use it as gaze primitives and calculate the average amplitudes of shifts in τ_{fix}, $z_t = \| \max(X_t) - \min(X_t) \| / \tau_{\text{fix}}$. Since feature z_t describes how much gaze points shift in a fixed interval, it can be regarded as an average motion speed of gaze shifts if τ_{fix} is small to some extent. Then, we learn several thresholds to classify the types based on samples of the motion speeds collected from gaze data. As a result of the classification, the types of gaze primitives can be associated with biological definitions of eye movement types: e.g., fixations, pursuits and saccades.

5.2.3 Top-down Modeling

In addition to the bottom-up approach based on the PSDM, we can also directly model the gap structures in a top-down manner. For this case, we first introduce

a parametric surface function that describes gap structures in a spatiotemporal patch. In addition, we extract salient points from the patch as samples to estimate parameters of the function. Then, Equation (5.2) can be interpreted as a problem of finding a surface that effectively discriminates positives from negatives. This approach can easily consider prior knowledge as to gap structures while we need to specify an appropriate function to represent them.

We particularly suppose the following two properties based on the observation of Figure 2.10 and Figure 2.11:

1. There are possible locations that a salient point exists in the spatiotemporal neighborhood around the the points of gaze, $\mathcal{N}(\boldsymbol{p})$, and thus the degree of gap occurrences has local extrema in $\mathcal{N}(\boldsymbol{p})$.

2. The variations of gap occurrences as well as the degree of saliency at the salient point along spatial and temporal directions are correlated. For example, salient points are possibly spatially distant from the points of gaze as they are temporally distant when salient regions are in motion. In addition, the degree of saliency mostly varies continuously as long as salient regions naturally change over time.

By taking them into account, we model the gap structure based on a quadratic function of the following form:

$$\begin{aligned} \boldsymbol{B}^{(\xi)}(\boldsymbol{p}) &= \boldsymbol{\beta}_{\xi}^{\mathrm{T}} \boldsymbol{w}_{\xi}(\boldsymbol{p}), \\ \boldsymbol{w}_{\xi}(\boldsymbol{p}) &= \mathrm{upperVec}\left(\left(d_{\mathrm{s}}, d_{\mathrm{t}}, L_{\mathcal{N}_{\xi}}(d_{\mathrm{s}}, d_{\mathrm{t}}) \right) \left(d_{\mathrm{s}}, d_{\mathrm{t}}, L_{\mathcal{N}_{\xi}}(d_{\mathrm{s}}, d_{\mathrm{t}}) \right)^{\mathrm{T}} \right), \end{aligned}$$

where $\boldsymbol{\beta}_{\xi}$ denotes coefficients of the function, $(d_{\mathrm{s}}, d_{\mathrm{t}}) = \arg\max_{x,y} L_{\mathcal{N}_{\xi}}(x, y)$ for given scale ξ, and $\mathrm{UpperVec}(\cdot)$ is an operator to obtain a flatten column vector of the upper triangular elements of matrices. Namely, we select salient points that follow the surface function from spatiotemporal patches which take the maximum degree of saliency. As for gaze-point prediction, $\boldsymbol{\beta}_{\xi}$ is a model parameter to be learned in a discriminant function, and $\boldsymbol{w}_{\xi}(\boldsymbol{p})$ serves as feature vector $\boldsymbol{\varphi}(\boldsymbol{p})$. Along with the bottom-up approach, we also concatenate the function of multiple scales to obtain $\boldsymbol{\varphi}(\boldsymbol{p})$ such as $\boldsymbol{\varphi}(\boldsymbol{p}) = (\boldsymbol{w}_1(\boldsymbol{p})^{\mathrm{T}}, \dots \boldsymbol{w}_{\xi_{\max}}(\boldsymbol{p})^{\mathrm{T}})^{\mathrm{T}}$ (and thus $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \dots, \boldsymbol{\beta}_{\xi_{\max}}^{\mathrm{T}})^{\mathrm{T}}$ since we introduce a linear function in Equation (5.1)).

The top-down modeling of structures appear in different applications: for example, object detection [Felzenszwalb et al., 2010] and face recognition

[Uřičář et al., 2012]. They learn the relative positions of object parts by introducing a quadratic function. Our model of gap structures is different from them in terms of describing not only spatial but temporal relationships and utilizing multiple-scale information.

5.3 Experiments

The experiments in this chapter are aimed at evaluating the effectiveness of the learned spatiotemporal correlations (modeled gap structures learned for each type of gaze primitives) with the task of gaze point prediction from videos. We adopted several combinations of public datasets and saliency maps to investigate if the effectiveness can be consistent regardless of videos and input saliency.

5.3.1 Datasets, Saliency Maps and Their Evaluations

Datasets and saliency maps

We introduced the following two datasets:

CRCNS-ORIG [Itti and Baldi, 2009] (CRCNS) [(i)] contains 50 videos with a variety of genres including surveillance videos, game plays, TV news and commercial films. Each video was watched by 4-6 subjects who were instructed to "follow the main actors and actions". The frame rate of videos is 30 fps.

ASCMN database [Riche et al., 2012] (ASCMN) [(ii)] contains 24 videos consisting of outdoor scenes, surveillance videos, videos of human crowds, etc. The videos in the database include parts of CRCNS-ORIG [Itti and Baldi, 2009], Vasconcelos's database [Mahadevan and Vasconcelos, 2010][(iii)] and a standard complex-background video surveillance database [Li et al., 2004][(iv)]. Parts of them contain objects with sudden motion events, which can possibly provide spatiotemporal gaps. Each video has 10 subjects who were not instructed particularly during experiments. The frame rate is 15 fps.

We also adopted the following three models of saliency maps:

[(i)] http://crcns.org/data-sets/eye/eye-1
[(ii)] http://www.tcts.fpms.ac.be/attention/?article38/saliency-benchmark
[(iii)] http://www.svcl.ucsd.edu/projects/background_subtraction/ucsdbgsub_dataset.htm
[(iv)] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

Itti's model [Itti et al., 1998] (IT) is one of traditional saliency maps, which now serves as a baseline in many studies. We chose color, intensity, orientation channels and did not adopt motion channels for the sake of fairness to the other models.

Cheng's model [Cheng et al., 2011] (RC) is a family of salient region detection techniques that extract a region with statistical irregularity in a given image. This model first segments images into small superpixels [Felzenszwalb and Huttenlocher, 2004], and give the degree of saliency to them based on the rarity of color.

Torralba's model in Judd et al. [Judd et al., 2009] (TR) is a simple saliency map based on the rarity of responses from various subband pyramids, which is utilized in [Judd et al., 2009].

Evaluation metrics

As a measure to evaluate gaze prediction maps, we introduce the normalized scanpath saliency (NSS) [Parkhurst et al., 2002]. The NSS evaluates the correlation between saliency maps (i.e., prediction results; gaze prediction maps in our study) and observed points of gaze (ground truths). First, an evaluation score for saliency map S with single point $\boldsymbol{p}$ is calculated as follows:

$$NSS(S, \boldsymbol{p}) = \frac{S(\boldsymbol{p}) - \mu(S)}{\sigma(S)}, \tag{5.4}$$

where $\mu(S)$ and $\sigma(S)$ are the mean and standard deviation of the degree of saliency in S, respectively. Then, given a set of gaze points at frame t, $\mathcal{X}_t = \{\boldsymbol{p}_t^{(n)} \mid n = 1, \ldots, N_t\}$ where N_t is the number of samples, we calculate Equation (5.4) for all the samples and average them to get an NSS score. When we deal with videos, that is, we have sequence of saliency maps $\mathcal{S} = (S_1, \ldots, S_T)$ and corresponding gaze point sets $(\mathcal{X}_1, \ldots, \mathcal{X}_T)$, we calculate the NSS scores for all the pairs of S_t and $\mathcal{X}_t$ and average them for the evaluation.

Preliminary experiments

We first quantitatively analyze how much spatiotemporal gaps exist between the points of gaze and salient regions in public datasets and how they are affected by the types of gaze primitives. To this end, we extend the NSS defined in Equa-

tion (5.4) by calculating the mean and standard deviation of saliency in local patches of several different scales with the center at $\boldsymbol{p}$. If NSSs at the points of gaze are high, subjects look at salient regions without gaps. Otherwise, there are higher salient regions around the gaze points and there are gaps between the points of gaze and salient regions that are possibly focused on.

As for the types of gaze primitives, we divided gaze data into four subsets where each of them corresponds to one primitive type from the four, $\mathcal{E} = \{e_1, \ldots, e_4\}$. Specifically, we calculate 25, 50 and 75 percentiles in the distributions of average motion speed z_t in Section 5.2.2 and utilize them as thresholds to classify the types of primitives. Note that we set τ_{fix} to 0.2 sec empirically and the ascending order of the amplitude was $e_1 < e_2 < e_3 < e_4$. We can regard them as fixations, slow pursuits, fast pursuits and saccades, respectively.

Table 5.1 demonstrates NSS scores in a variety of combinatorial conditions. We resized images into 80×60 pixel resolution, and used patches of $(\delta_{\text{x}}, \delta_{\text{y}}) = (5,5), (15,15)$ to calculate NSSs (i.e., 11×11 and 31×31 pixel patches; the same settings as Section 2.4.4). We found that NSSs tended to decrease as motion speeds increased (from e_1 to e_4), which indicated that there were more spatiotemporal gaps when gaze points moved larger.

Another finding is that the NSSs tended to decrease as the sizes of patches got smaller. It demonstrates that gaze points tend to be directed globally salient but locally non-salient regions. We can take into account of such characteristics by introducing multi-scale local scene structures presented in Section 5.2.2.

5.3.2 Parameter Settings

In the gap structure modeling, parameters to be trained are as follows:

- Neighborhoods $\mathcal{N}_1(\boldsymbol{p}) \ldots \mathcal{N}_{\xi_{\max}}(\boldsymbol{p})$ and the number of scales, $\xi_{\max}$.
- Codebooks of saliency primitive types $\mathcal{M}_1, \ldots, \mathcal{M}_{\xi_{\max}}$ and their sizes $N_1, \ldots, N_{\xi_{\max}}$.
- $N_{\text{etype}} - 1$ thresholds to give the types of gaze primitives as well as N_{etype}.
- Model parameters $\boldsymbol{\beta}_{e_1}, \ldots, \boldsymbol{\beta}_{e_{N_{\text{etype}}}}$ for N_{etype} types of gaze primitives.

$\xi_{\max}$ was empirically defined as $\xi_{\max} = 2$ and the spatial sizes of $\mathcal{N}_1$ and $\mathcal{N}_2$, i.e., $(\delta_{\text{x}}, \delta_{\text{y}})$ were defined as $(5,5)$ and $(15,15)$ in 80×60 pixel-frames, respectively. The temporal sizes of $\mathcal{N}_1$ and $\mathcal{N}_2$, i.e., δ_{t} were both 0.4 sec. The number of gaze

Table 5.1: NSS scores in various combinatorial conditions. Original NSS is calculated in a whole image of 80×60 pixels.

Datasets	saliency maps (original NSS)	Types	31×31	11×11
CRCNS	IT (0.75)	e_1	0.51	0.14
		e_2	0.44	0.11
		e_3	0.34	0.07
		e_4	0.25	0.04
	RC (0.91)	e_1	0.59	0.16
		e_2	0.52	0.14
		e_3	0.42	0.11
		e_4	0.34	0.09
	TR (0.73)	e_1	0.6	0.2
		e_2	0.53	0.16
		e_3	0.48	0.14
		e_4	0.39	0.12
ASCMN	IT (0.62)	e_1	0.38	0.09
		e_2	0.31	0.05
		e_3	0.21	0.01
		e_4	0.17	0.01
	RC (0.59)	e_1	0.34	0.09
		e_2	0.26	0.05
		e_3	0.21	0.04
		e_4	0.18	0.03
	TR (0.39)	e_1	0.29	0.08
		e_2	0.24	0.05
		e_3	0.15	0.02
		e_4	0.14	0.02

primitive types were set as the same as Section 5.3.1, $N_{\text{etype}} = 4$. On the other hand, $\mathcal{M}_1, \mathcal{M}_2, N_1, N_2, \boldsymbol{\beta}_{e_1}, \ldots, \boldsymbol{\beta}_{e_{N_{\text{etype}}}}$ and $N_{\text{etype}} - 1$ thresholds were estimated in a training dataset. $N_{\text{etype}} - 1$ thresholds were given as 25, 50, and 75 percentiles in a training dataset.

5.3.3 Evaluation Scheme

In order to evaluate a generalization capability on videos, we conducted a leave-one-out scheme by splitting data based on video IDs (and that is, we did not distinguish subjects). Specifically, we first divided a dataset consisting of V videos into $V - 1$ training videos and 1 test video. From a training subset, we collected

positive samples from a set of points where subjects looked. As for negatives, we randomly selected samples of the same size as positives from videos. Then, we trained parameters so as to get the highest area-under-the-curve (AUC) score of a receiver operating characteristic curve with false-positive vs. true-positive rates. We here adopted a Fisher's discriminant analysis so as to conduct evaluations with a simple learning technique. With a trained model, we evaluated the degree of gaze-point existence (Equation (5.1) or Equation (5.3)) for all of the frames in the test video and generate a sequence of gaze-prediction maps. We finally calculated an NSS score by averaging the NSSs in Equation (5.4) over all the pairs of obtained prediction maps and corresponding gaze data.

In the experiments, we evaluated (1) which of bottom-up/top-down modelings of gap structures was effective and (2) whether gaze primitive types contributed to the prediction or not. Specifically, we tested methods defined by the combinations of modeling approaches (Section 5.2.2 and Section 5.2.3) and prediction formulae (Equation (5.1) and Equation (5.3)). In what follows, we refer to them as $\mathbf{M}_{\mathrm{BU}}$ (bottom-up + Equation (5.1)), $\mathbf{M}_{\mathrm{BU+E}}$ (bottom-up + Equation (5.3)), $\mathbf{M}_{\mathrm{TD}}$ (top-down + Equation (5.1)), $\mathbf{M}_{\mathrm{TD+E}}$ (top-down + Equation (5.3)). As a baseline method, we adopted original saliency maps $\mathbf{M}_{\mathrm{ORIG}}$ (i.e., IT, RC and TR introduced in Section 5.3.1) as well as a method that modified [Riche et al., 2012] $\mathbf{M}_{\mathrm{SM}}$ for the sake of fairness. The original method in [Riche et al., 2012] basically utilized broadened saliency maps to fill spatiotemporal gaps. In our experiments, baseline method M_{SM} followed this idea and smoothed saliency maps, where the smoothing parameter was learned so as to get the highest AUC score in a training subset. In addition, we regarded the degree of smoothed saliency as a feature value for each pixel, and learned it in a discriminant function along with the proposed method. **SM** can be extended by training models with respect to each type of gaze primitives and average them over the types ($\mathbf{M}_{\mathrm{SM+E}}$). Consequently, we evaluated $\mathrm{M}_{\mathrm{ORIG}}$, M_{SM}, $\mathrm{M}_{\mathrm{SM+E}}$, M_{TD}, $\mathrm{M}_{\mathrm{TD+E}}$, M_{BU}, $\mathrm{M}_{\mathrm{BU+E}}$ under 2 dataset $\times$ 3 saliency map conditions.

5.3.4 Results and Discussions

Table 5.2 shows NSS scores for all the conditions. These results demonstrated the effectiveness of M_{BU}, $\mathrm{M}_{\mathrm{BU+E}}$, M_{TD}, $\mathrm{M}_{\mathrm{TD+E}}$ compared to the baseline methods ($\mathrm{M}_{\mathrm{ORIG}}$, M_{SM}, $\mathrm{M}_{\mathrm{SM+E}}$). Although the NSS scores of the baseline methods had a variation with regard to the saliency maps, the scores of our methods were very competitive. It indicates the independence of our models to input saliency maps

Table 5.2: Average NSS scores over videos.

		M_{ORIG}	M_{SM}	M_{SM+E}	M_{TD}	M_{TD+E}	M_{BU}	M_{BU+E}
CRCNS	IT	0.752	0.859	0.847	0.979	0.980	1.135	**1.208**
	RC	0.927	1.002	1.021	1.034	1.035	1.152	**1.212**
	TR	0.742	0.858	0.886	1.033	1.034	1.100	**1.152**
ASCMN	IT	0.623	0.745	0.741	0.821	0.820	0.876	**0.900**
	RC	0.603	0.659	0.651	0.714	0.715	0.765	**0.775**
	TR	0.388	0.465	0.466	0.718	0.719	0.774	**0.817**

to describe gap structures. Comparing methods with or without the consideration of gaze primitive types, the bottom-up approach M_{BU+E} only performed improvements from M_{BU} while other methods show slight changes.

Figure 5.3 depicts qualitative results of gaze-prediction maps and NSS scores. These results demonstrate the following features of the proposed method:

Proposed method vs. baseline methods In Examples (A) and (E), a car was running out of the frame, and most of subjects were trying to pursue it. In addition, Example (D) shows the situation where subjects pursue the player running to left. Obviously there are gaps between the points of gaze and the targets in both examples, and thus baseline methods providing a large degree of gaze-point existence at the targets get low NSS scores. On the other hand, the proposed method incorporates such gaps into prediction and provide a large degree of gaze-point existence where the points of gaze exist and succeeded in significantly improving the NSS scores.

Differences in saliency maps Examples (B) and (C) depict the comparison of different saliency maps, IT and RC. Since RC looks for small superpixels that contain a rare color, the baseline methods show high responses at the black regions in the top-left of a frame. That brings the significant differences in NSS scores not only in the baseline methods but in the proposed method, although averaged scores in Table 5.2 show small differences among saliency maps in the proposed method.

Failure cases In Example (F), the original saliency map (TR) was able to capture the points of gaze precisely. Even such cases, the proposed method tries to consider a spatiotemporal gap since there are many samples with gaps in training datasets, which sometimes provides large degree of gaze-point existence at inappropriate locations and decreases a score.

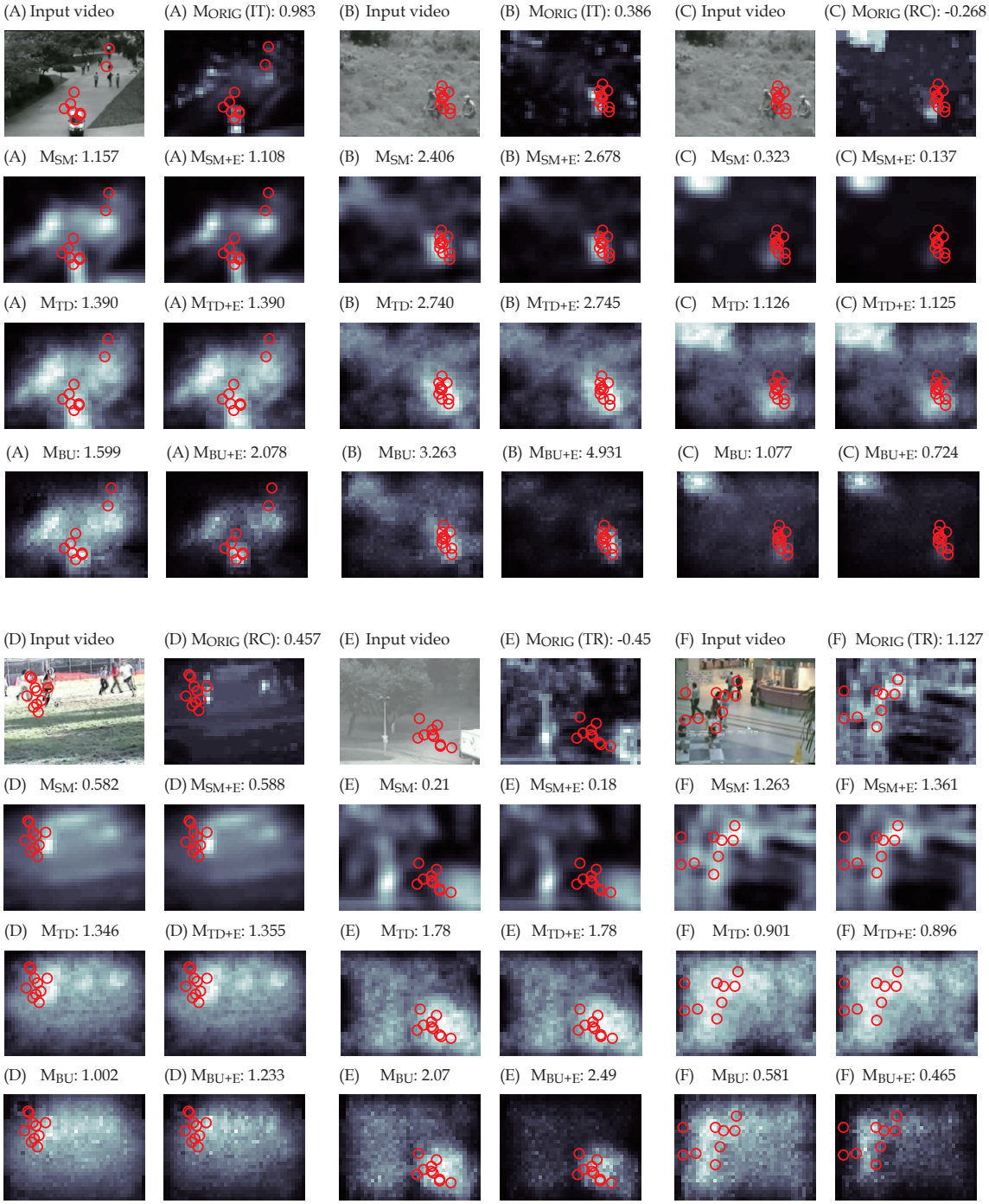

Figure 5.3: Qualitative results and corresponding NSS scores averaged over subjects in a frame. Luminance indicates the degree of gaze-point existence. Red points indicate a set of gaze points, where each point corresponds to an individual subject in [Riche et al., 2012]. Parts of the photos in this figure are contained in the dataset provided by [Li et al., 2004, Mahadevan and Vasconcelos, 2010, Itti and Baldi, 2009].

Figure 5.4 visualizes the gap structures that well discriminated positive samples (points of gaze) from negatives (random points) in the bottom-up approach based on the PSDM. We reconstructed the structures by giving coefficients of the trained discriminant function as activations to specific types of saliency primitives (higher values in coefficients contribute to a larger degree of gaze-point existence), and summing up the activated primitives. In the smaller scale, salient regions tend to be the past of the gaze points, where the size of delays depended on the types of gaze primitives and saliency maps. Meanwhile for the larger scale, there are salient regions after the points of gaze for most of the cases. It indicates that subjects were somewhat predictive to salient regions, but could not accurately follow them without a delay. In addition, these results also indicate the effectiveness of bottom-up approaches compared to the top-down ones since several structures are hard to represent by a quadratic function.

Comparing $\mathrm{M}_{\mathrm{BU+E}}$ with M_{BU}, highlighted regions indicating a large degree of gaze-point existence are more sparse in $\mathrm{M}_{\mathrm{BU+E}}$ as shown in Examples (A), (B), (E), (F) in Figure 5.3. We can observe such results when one of the prediction scores for a certain type of gaze primitives is particularly high. In other words, other methods including $\mathrm{M}_{\mathrm{TD+E}}$ failed to involve such differences between gaze primitive types. With regard to Example (E), Figure 5.5 visualizes each of model outputs by the difference of color. In the 3rd row of the figure, we gave each pixel a 3-d value $(B_{e_1}(\boldsymbol{p}), 0.5(B_{e_2}(\boldsymbol{p}) + B_{e_3}(\boldsymbol{p})), B_{e_4}(\boldsymbol{p}))$ in an RGB order where each of them roughly corresponds to fixations, pursuits and saccades. When there was a target in motion, model outputs of pursuits (green) became much higher than the others, which made final outputs more sparse. In addition, there was also a small probability of observing saccades (blue). For example, saccades can be found when trying to attend the target (points at the left side of the frame in the 3rd column) or escaping from the target (those at the bottom-right in the 4th column).

Figure 5.6 presents experimental results with artificial stimuli recorded in the CRCNS-ORIG. Note that these prediction maps are the result of the learning from the other (more natural) videos in the dataset. In the example, a red blob at the top-left of frames disappeared and at the same time a pink blob appeared at the bottom-right at 8/30 sec. After 10/30 sec, all the subjects started attending the pink blob at 18/30 sec. Among the methods, $\mathrm{M}_{\mathrm{SM+E}}$ and $\mathrm{M}_{\mathrm{TD+E}}$ incorrectly predict the existence of gaze points at the bottom-right before the appearances of the pink blob. These prediction results seem to happen because the methods learn anticipatory gaze motions from training data, which did not happen when objects

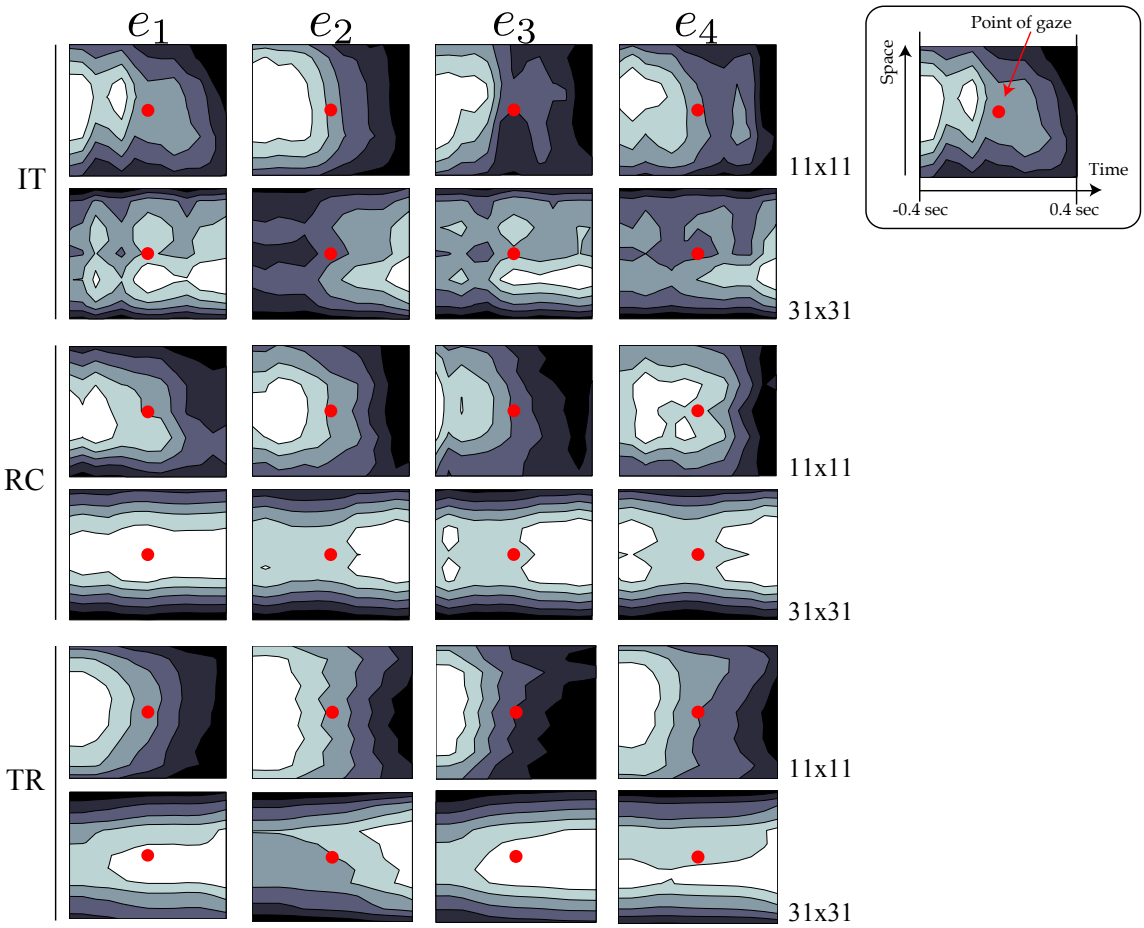

Figure 5.4: Gap structures contributing to the prediction for the combination of saliency maps and the types of gaze primitives. The red point in the center of the contour figures show the locations of gaze points. This figure is a part of author's publication [Yonetani et al., 2013] copyrighted by Association for Computing Machinery.

appeared suddenly like this example. Just after the appearance of the new blob, NSS scores decrease first in M_{ORIG} and then in M_{SM+E} since they cannot consider a reaction delay correctly. On the other hand, M_{TD+E} and M_{BU+E} predict the existence of gaze points at the top-left and keep the NSS scores higher. In particular, the shift of the maxima in gaze prediction maps by M_{BU+E} is fairly accurate as shown in the results from 8/30 to 18/30 sec. These results support the appropriateness of the bottom-up description of spatiotemporal correlations in terms of describing gaze behavior with expected reaction delays accurately. Only, the baseline methods M_{ORIG} and M_{SM+E} obtained a higher NSS when there existed no gaps obviously such as 0/30 to 6/30 sec.

Finally, this study introduced a simple assumption for gaze primitive types,

that is, each type can appear with equal probability, independently and identically for spatial and temporal directions. The prior probability on the types of gaze primitives can be biased; for example, the saccadic primitives can be less observed than other types. In addition, gaze primitive types at a certain spatiotemporal point can be statistically conditioned by those at its spatiotemporal neighborhood. Currently, the degree of improvements in NSSs from M_{BU} to M_{BU+E} is smaller than that from M_{ORIG} to M_{BU}. Then, there is still room for further improvements by considering the aspect above. A promising approach is to introduce state-space models such as [Pang et al., 2008]. It assumes a Markov property for occurrences of gaze primitive types and gaze positions. By taking this into account, we can dynamically select models to be used based on the gaze primitive types which are likely to appear.

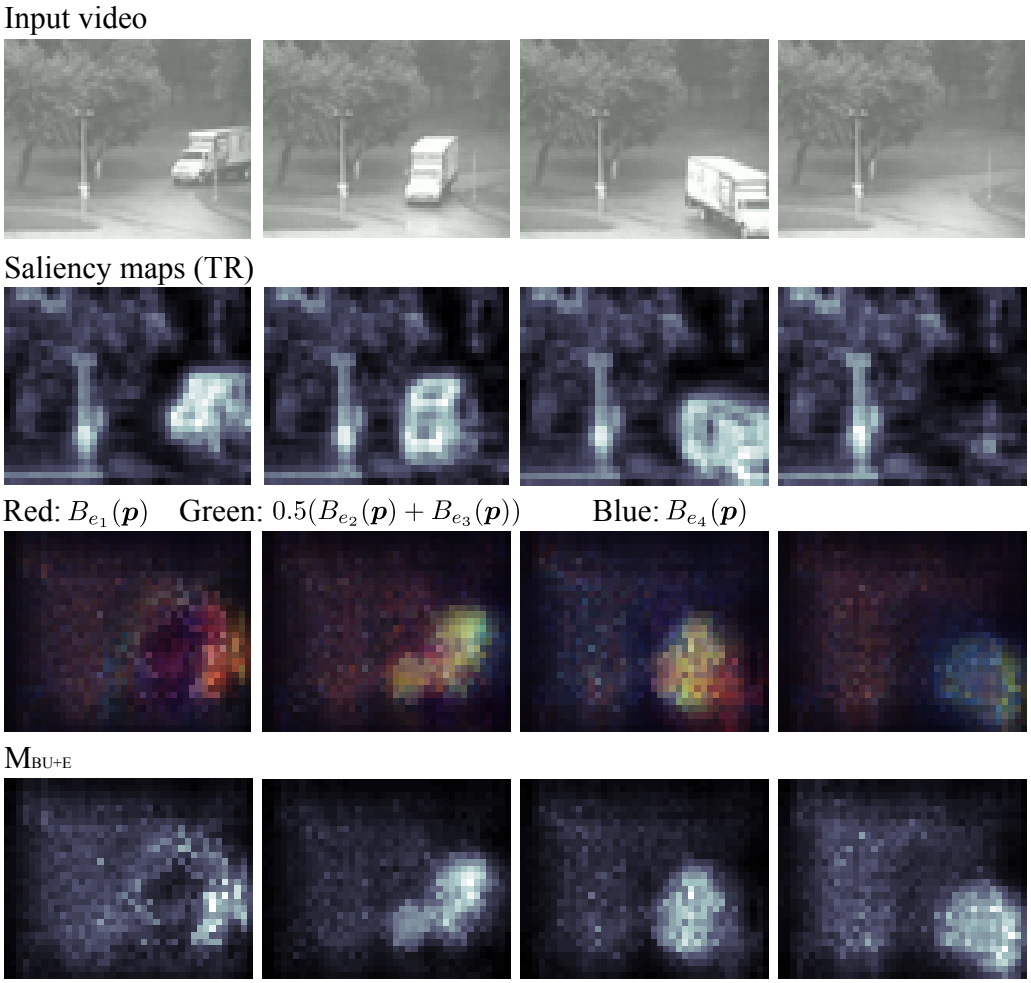

Figure 5.5: Differences in outputs of the bottom-up approach trained for each type of gaze primitives. In the 3rd row, red points show a large degree of gaze-point existence for $B_{e_1}(\boldsymbol{p})$, green for $0.5(B_{e_2}(\boldsymbol{p}) + B_{e_3}(\boldsymbol{p}))$ and blue for $B_{e_4}(\boldsymbol{p})$. Parts of the photos in this figure are contained in the dataset provided by [Mahadevan and Vasconcelos, 2010]. This figure is a part of author's publication [Yonetani et al., 2013] copyrighted by Association for Computing Machinery.

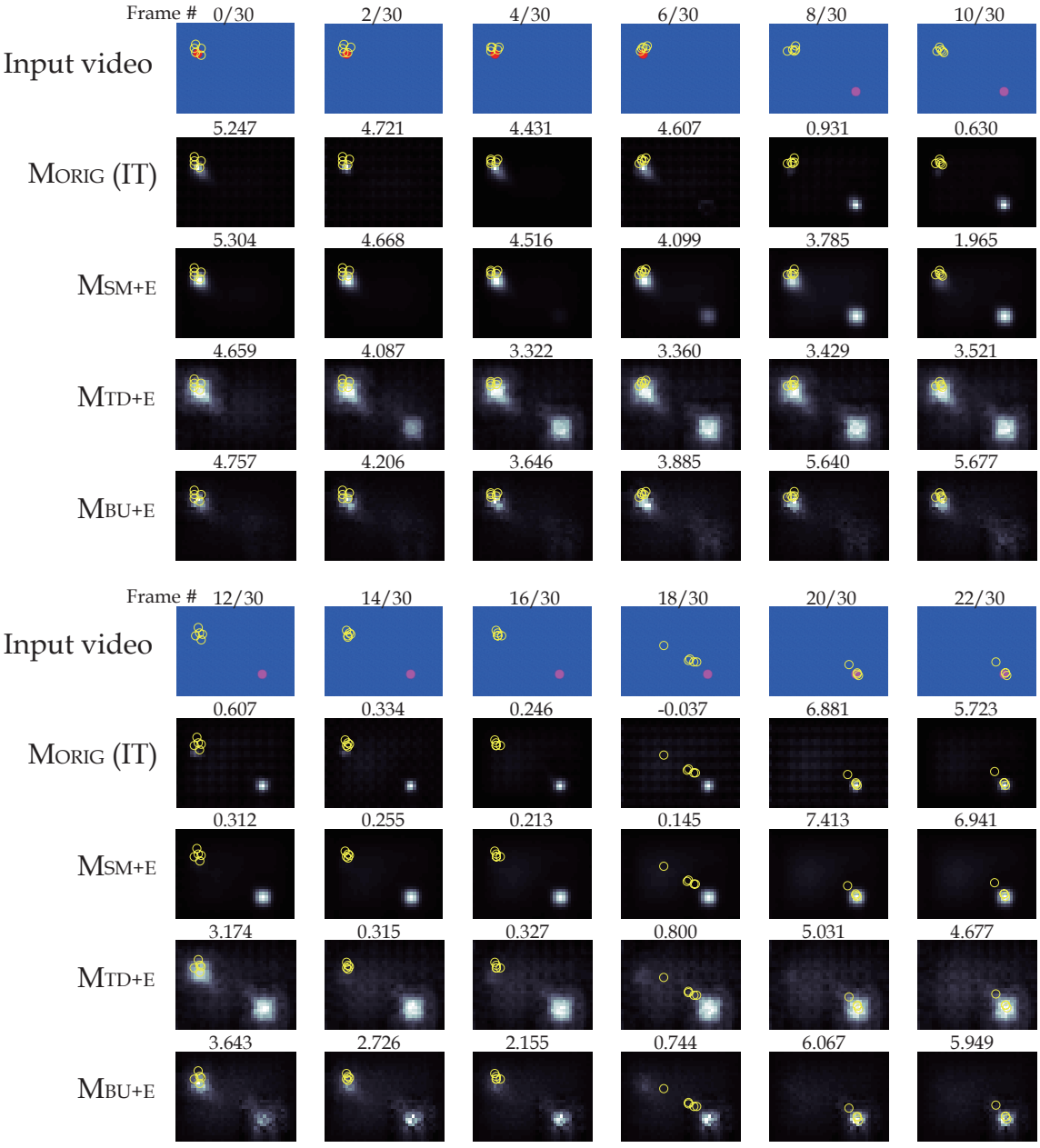

Figure 5.6: Qualitative results with artificial stimuli. Luminance indicates the degree of gaze-point existence. Yellow points indicate a set of gaze points, where each point corresponds to an individual subject in [Itti and Baldi, 2009]. Parts of the photos in this figure are contained in the dataset provided by [Itti and Baldi, 2009].

Chapter 6

Conclusions

6.1 Summary

In this thesis, we presented a novel framework to describe the spatiotemporal correlations between video and gaze data. We particularly covered a situation where human observers were watching a variety of videos taken in real environments. In such situations, videos contain various visual events and time-varying scene structures. Their influences upon gaze dynamics were characterized by the relationships named event-level spatiotemporal gaps and scene-level correlations, respectively. In order to describe these twofold relationships, we proposed a series of saliency dynamics models in Chapter 2 that described visual events and scene structures using saliency primitives.

In the following chapters, we introduced specific descriptions of spatiotemporal correlations based on our framework and evaluated their effectiveness via several gaze behavior analyses in real environments. First, we introduced a description of event-level spatiotemporal gaps under a simplified situation where scene structures were constant and visual events were given in Chapter 3. Specifically, we calculated the temporal distances between saliency and gaze primitives and leveraged them for the task of attentional target identification, which we referred to as the Gaze Probing. We demonstrated that the Gaze Probing was able to capture the temporal synchronizations between visual events and gaze reactions well and realize the robust identification to gaze tracking errors.

Chapter 4 covered intentionally-designed videos with various visual events including frequent scene changes. We analyzed how scene-level correlations were differently characterized depending on the time-varying types of scene structures. The saliency primitives learned by the object-based saliency dynamics

model were classified into several types to describe the types of scene structures as well as to characterize the scene-level correlations. We utilized the correlation information for an attentive state estimation task and confirmed the effectiveness of proposed description.

Finally, Chapter 5 addressed overall spatiotemporal correlations by analyzing how the event-level spatiotemporal gaps were influenced by the scene-level correlations. To this end, we introduced a model of gap structures that jointly described spatiotemporal gaps and scene structures based on the patch-based saliency dynamics model. By statistically learning the modeled gap structures with respect to each type of gaze primitives, we involved the scene-level correlations between scene structures and gaze dynamics. The proposed description was evaluated via gaze point prediction from videos including unedited natural ones. Experimental results demonstrated that the proposed method was able to predict gaze points more accurately than traditional saliency maps.

6.2 Future Work

We should explicitly mention again that we focused on the description of spatiotemporal correlations that were able to be observed via visual content analyses and gaze tracking. Although cognitive and neurological reasoning is invisible in our framework, we are sure the proposed framework can contribute to the analyses of various gaze behavior in real environments including but not limited to the situation where humans watch videos. The following section discusses some limitations and possible extensions of our framework as well as some applications to other gaze behavior than watching videos.

6.2.1 Limitations and Extensions of Saliency Dynamics Models

(a) Semantic extension of saliency dynamics models

The visual content analyses adopted in this thesis mainly focused on the physical aspects (i.e., saliency) of visual events, and semantic aspects such as categories of objects, scenes and actions were mostly invisible as mentioned in Section 4.5.3. If we try to handle gaze behavior during specific tasks such as “choosing favorite objects from movies” and “trying to understand and summarize what was explained in TV news”, modeling of saliency dynamics without semantic aspects are not always sufficient for gaze behavior analyses.

As briefly reviewed in Section 2.1.2, the introduction of semantic aspects into saliency models has become a popular issue recently. One of the basic approaches is to incorporate object detection scores as a feature [Cerf et al., 2007, Judd et al., 2009, Borji, 2012]. In addition, current trends include hiring various machine learning techniques to learn saliency with semantic information in a data-driven manner [Wang et al., 2012, Goferman et al., 2012, Sharma et al., 2012, Mai et al., 2013]. By incorporating such saliency calculations into our saliency dynamics models, we can deal with visual events which are not salient in a bottom-up sense but conspicuous in a top-down sense due to the semantics.

Another relevant topic is to analyze semantic relations within a content, namely, the studies on the semantic aspects of visual events indicating "why they attract attention", besides our framework. We have recently proposed a model of the semantic relations with spatial layouts (i.e., semantic scene structures) for designed static contents (e.g., pictorial books) [Ishikawa et al., 2012]. The proposed model adopts a graph structure to represent a hierarchy from object instances to their categorical groups where the instances have spatial regions as a property. It enables us to describe various gaze actions such as "comparing several categories" and "examining various items of a specific category" by referring to transitions of the categories of interest from the sequences of objects being looked at. One of our future work is to extend this model so as to describe time-varying semantic scene structures. Since we introduce a graph structure for the representation of static contents in [Ishikawa et al., 2012], we need to deal with the dynamic changes of graphs, which has been well studied in the field of data mining [Sun et al., 2007, Yang et al., 2011].

(b) Multimodal extension with acoustic information

Videos generally contain not only visual but acoustic information, and they are associated with each other for many cases. In the light of gaze behavior analyses, both visual and acoustic information can attract our visual attention.

Fusing audiovisual information is one of the central issues in HCI and multimedia research fields. For example, speaker detection is a problem of finding a speaker from other non-speakers displayed together in a video. To address this problem, existing studies associate lip motions with audio signals such as [Pavlovic et al., 2000, Horii et al., 2008]. As a similar problem, studies on audio localization calculate correlations between visual and audio signals to extract local regions that possibly emit a sound [Kidron et al., 2007, Liu and Sato, 2009].

Regardless of many advances in the audiovisual fusion studies, it is still a difficult problem to incorporate acoustic information into saliency calculations. So far, [Ma et al., 2005] has introduced a model that enhanced all of the saliency in a scene by the acoustic information. In addition, [Schauerte and Stiefelhagen, 2013] has recently proposed a model of acoustic saliency based on the Bayesian surprise model [Itti and Baldi, 2009]. A next step along this issue is to associate the acoustic saliency with local visual events in scene structures to enhance their saliency.

6.2.2 From Action-Reaction to Interaction

(c) Mental state estimation for various assistances

In Chapter 4, we estimated attentive states as one of mental states based on the framework. When extending our framework to interactive situations, estimating mental states from eyes is one thing, but giving a feedback to observers is another. Mental states play a crucial role when interactive systems try to display suitable information to observers in a timely manner. For example, Info-concierge is a system of proactively interacting with users based on the interests of users estimated from their gaze behavior [Hirayama et al., 2011].

There are several issues when incorporating our framework into interactive systems. First, we need a realtime algorithm to estimate mental states to give feedbacks in a timely manner. With regard to our attentive state estimation proposed in Chapter 4, it is not difficult to estimate attentive states near realtime essentially if saliency dynamics models have already been applied to displayed videos and estimation models have been trained in relevant types of scene structures before their playback.

The second issue is how to give feedbacks — when, where and what to display. To determine a timing of feedbacks, attentive states can be a crucial clue since observers can accept new information when they are in lower level of attentiveness. In addition, several studies have proposed a method to infer an appropriate timing to interact with users based on the mental states such as interruptibility [Fogarty et al., 2005] and engagement [Nakano and Ishii, 2010]. The determination of the places to put feedbacks in a screen is a more difficult problem. The feedbacks should be displayed at the places where observers can easily attend from current gaze locations and where informative regions in original contents are not occluded, although such gaze and scene structures can change over time. Several studies on augmented reality have introduced a system to dis-

play texts at suitable locations [Leykin and Tuceryan, 2004, Gabbard et al., 2005, Orlosky et al., 2013]. With regard to what to display, we are currently studying gaze-based recommender systems based on mental states estimated from gaze like [Yoshitaka et al., 2007]. Our approach tries to model mental states of observers by the mixtures of several aspects of interests, e.g., "looking for healthy foods" and "looking for cute animals", where the aspects are associated with visual attributes of displayed items [Shimonishi et al., 2013]. One of our future work is to integrate these techniques altogether to construct interactive systems that can behave adaptively to users.

(d) Cooperative analysis and understanding of multiple first person views

Analyzing human gaze behavior via first person view (FPV) videos has obtained a lot of attention in the field of computer vision recently. In particular, several studies have focused on the measurement of social gaze behavior when several interaction participants moved and talked with each other [Park et al., 2012, Fathi et al., 2012].

Incorporating our framework into the analyses of interactions by multiple participants is another extension from action-reaction to interaction. Let us suppose a situation where two participants (let's say A and B) have an interaction. For participant A, our framework can analyze what types of actions (visual events) provided by the other participant B in the FPV can attract A's attention. As presented in [Tsotsos, 2011], observed characteristics resulted from overt attention are not only eye movements but head and body movements. Thus, we need to analyze gaze dynamics as well as head motions observed as global motions (i.e., camera motions) and body motions partially observed in the FPV as reactions to B's actions. To this end, the first step to the extension is motion segmentation such as [Rath and Makur, 1999, Chan and Vasconcelos, 2008, Zhang et al., 2013]. The obtained global motions can contribute to not only the understanding of how participants react to visual events by their whole body but to the enhancement of saliency maps based on the influences of egocentric motions upon visual attention presented in [Yamada et al., 2010].

Another important aspect in the aforementioned example is that B's actions can sometimes perform as reactions resulted from A's actions in an interactive scene. Thus, we need to introduce our framework for each of the participants and couple them to conduct a cooperative analysis. That is, we analyze how B's actions affect A's reactions, and in turn how the reactions af-

fect next B's actions. Several studies have aimed to discover characteristic action-reaction patterns from multi-party interactions in a data-mining fashion [Jayagopi and Gatica-Perez, 2010, Jayagopi et al., 2012, Yu et al., 2012]. These studies basically extract order relations and co-occurrences of specific behavior as a motif of interactions. On the other hand, the extension of our framework is different from them in terms of utilizing event-level spatiotemporal gaps between actions and reactions influenced by whole participants' behavior, i.e., scene-level correlations. It enables us to describe novel interaction patterns such as "B reacts to A's hand gestures with a large reaction delay since B is in the lower level of attentiveness". Mining such patterns will lead to the deeper understanding of multi-party interactions in future work.

Bibliography

[Achanta et al., 2008] Achanta, R., Estrada, F., Wils, P., and Süsstrunk, S. (2008). Salient Region Detection and Segmentation. In *Proceeding of the International Conference on Computer Vision Systems*, pages 66–75.

[Baldi and Itti, 2010] Baldi, P. and Itti, L. (2010). Of Bits and Wows: A Bayesian Theory of Surprise with Applications to Attention. *Neural Networks*, 23(5):649–666.

[Becker and Fuchs, 1985] Becker, W. and Fuchs, A. F. (1985). Prediction in the Oculomotor System: Smooth Pursuit during Transient Disappearance of a Visual Target. *Experimental Brain Research*, 57:562–575.

[Bednarik et al., 2012] Bednarik, R., Vrzakova, H., and Hradis, M. (2012). What Do You Want to Do Next : A Novel Approach for Intent Prediction in Gaze-based Interaction. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, pages 83–90.

[Beymer and Flickner, 2003] Beymer, D. and Flickner, M. (2003). Eye Gaze Tracking Using an Active Stereo Head. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2(2):451–458.

[Bishop, 2006] Bishop, C. M. (2006). 9. Mixture Models and EM. In *Pattern Recognition and Machine Learning (Information Science and Statistics)*, pages 423–460. Springer-Verlag New York, Inc.

[Borji, 2012] Borji, A. (2012). Boosting Bottom-up and Top-down Visual Features for Saliency Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Borji and Itti, 2012] Borji, A. and Itti, L. (2012). State-of-the-art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

[Borji et al., 2012] Borji, A., Sihite, D., and Itti, L. (2012). Probabilistic Learning of Task-specific Visual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Brandherm et al., 2007] Brandherm, B., Prendinger, H., and Ishizuka, M. (2007). Interest Estimation Based on Dynamic Bayesian Networks for Visual Attentive Presentation Agents. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 346–349.

[Bregler, 1997] Bregler, C. (1997). Learning and Recognizing Human Dynamics in Video Sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–574.

[Brezeale and Cook, 2008] Brezeale, D. and Cook, D. J. (2008). Automatic Video Classification: A Survey of the Literature. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 38(3):416–430.

[Brox et al., 2004] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. (2004). High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *Proceedings of the European Conference on Computer Vision*, pages 25–36.

[Bruce and Tsotsos, 2009] Bruce, N. and Tsotsos, J. (2009). Spatiotemporal Saliency : Towards a Hierarchical Representation of Visual Saliency. In *Proceedings of the International Workshop on Attention in Cognitive Systems*.

[Cerf et al., 2007] Cerf, M., Harel, J., Einhäuser, W., and Koch, C. (2007). Predicting Human Gaze Using Low-Level Saliency Combined with Face Detection. *Proceedings of the Conference on Neural Information Processing Systems*, pages 1–8.

[Chan and Vasconcelos, 2008] Chan, A. and Vasconcelos, N. (2008). Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926.

[Chaudhry et al., 2009] Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on Nonlinear Dynamical Systems for the Recognition of Human Actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1932–1939.

[Chen et al., 2008] Chen, J., Tong, Y., Gray, W., and Ji, Q. (2008). A Robust 3D Eye Gaze Tracking System Using Noise Reduction. *Proceedings of the Symposium on Eye tracking research & applications*, pages 189–196.

[Cheng et al., 2011] Cheng, M., Zhang, G., Mitra, N., Huang, X., and Hu, S. (2011). Global Contrast Based Salient Region Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Cootes et al., 2001] Cootes, T., Edwards, G., and Taylor, C. (2001). Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

[Cotsaces et al., 2006] Cotsaces, C., Nikolaidis, N., and Pitas, I. (2006). Video Shot Detection and Condensed Representation: A Review. *IEEE Signal Processing Magazine*, 23(2):28–37.

[Cremers, 2006] Cremers, D. (2006). Dynamical Statistical Shape Priors for Level Set-based Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1262–1273.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893.

[Dollar et al., 2005] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior Recognition via Sparse Spatio-temporal Features. In *Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72.

[Doretto et al., 2003] Doretto, G., Chiuso, A., Wu, Y.-N., and Soatto, S. (2003). Dynamic Textures. *International Journal of Computer Vision*, 51(2):91–109.

[Dorr et al., 2010] Dorr, M., Martinetz, T., Gegenfurtner, K., and Barth, E. (2010). Variability of Eye Movements When Viewing Dynamic Natural Scenes. *Journal of Vision*, 10(10):1–17.

[Eckstein, 2011] Eckstein, M. P. (2011). Visual Search: A Retrospective. *Journal of Vision*, 11(5).

[Eivazi and Bednarik, 2011] Eivazi, S. and Bednarik, R. (2011). Predicting Problem-solving Behavior and Performance Levels from Visual Attention Data.

In *Proceedings of the Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI*, pages 9–16.

[Eivazi et al., 2012] Eivazi, S., Bednarik, R., Tukiainen, M., von und zu Fraunberg, M., Leinonen, V., and Jääskeläinen, J. (2012). Gaze Behaviour of Expert and Novice Microneurosurgeons Differs during Observations of Tumor Removal Recordings. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*.

[Fathi et al., 2012] Fathi, A., Hodgins, J. K., and Rehg, J. M. (2012). Social Interactions: A First-person Perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Fei-Fei and Perona, 2005] Fei-Fei, L. and Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531.

[Felzenszwalb and Huttenlocher, 2004] Felzenszwalb, P. and Huttenlocher, D. (2004). Efficient Graph-based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181.

[Felzenszwalb et al., 2010] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

[Fogarty et al., 2005] Fogarty, J., Hudson, S. E., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. C., and Yang, J. (2005). Predicting Human Interruptibility with Sensors. *ACM Transactions on Computer-Human Interaction*, 12(1):119–146.

[Frintrop et al., 2005] Frintrop, S., Backer, G., and Rome, E. (2005). Goal-Directed Search with a Top-Down Modulated Computational Attention System. In Kropatsch, W. G., Sablatnig, R., and Hanbury, A., editors, *Pattern Recognition*, volume 3663 of *Lecture Notes in Computer Science*, pages 117–124. Springer Berlin Heidelberg.

[Gabbard et al., 2005] Gabbard, J., Swan, J., Hix, D., Schulman, R., Lucas, J., and Gupta, D. (2005). An Empirical User-based Study of Text Drawing Styles and

Outdoor Background Textures for Augmented Reality. In *Proceedings of the IEEE Virtual Reality*, pages 11–18.

[Gao and Vasconcelos, 2009] Gao, D. and Vasconcelos, N. (2009). Decision-theoretic Saliency: Computational Principles, Biological Plausibility, and Implications for Neurophysiology and Psychophysics. *Neural Computation*, 21:239–271.

[Goferman et al., 2012] Goferman, S., Zelnik-Manor, L., and Tal, A. (2012). Context-Aware Saliency Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926.

[Goldstein et al., 2007] Goldstein, R., Woods, R., and Peli, E. (2007). Where People Look When Watching Movies: Do All Viewers Look at the Same Place? *Computers in Biology and Medicine*, 37(7):957–64.

[Harada et al., 2011] Harada, T., Ushiku, Y., Yamashita, Y., and Kuniyoshi, Y. (2011). Discriminative Spatial Pyramid. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1617–1624.

[Harel et al., 2007] Harel, J., Koch, C., and Perona, P. (2007). Graph-based Visual Saliency. In *Proceedings of the Conference on Neural Information Processing Systems*, volume 19, pages 545–552.

[Hennessey et al., 2006] Hennessey, C., Noureddin, B., and Lawrence, P. (2006). A Single Camera Eye-gaze Tracking System with Free Head Motion. *Proceedings of the Symposium on Eye tracking research & applications*, pages 87–94.

[Hirayama et al., 2010] Hirayama, T., Dodane, J. B., Kawashima, H., and Matsuyama, T. (2010). Estimates of User Interest Using Timing Structures between Proactive Content-display Updates and Eye Movements. *IEICE Transactions on Information and Systems*, E-93D(6):1470–1478.

[Hirayama et al., 2011] Hirayama, T., Sumi, Y., Kawahara, T., and Matsuyama, T. (2011). Info-concierge: Proactive Multi-modal Interaction through Mind Probing. In *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference*.

[Hoffman, 1998] Hoffman, J. (1998). Visual Attention and Eye Movements. In Pashler, H., editor, *Attention*, volume 31, chapter 3, pages 119–153. Psychology Press.

[Horii et al., 2008] Horii, Y., Kawashima, H., and Matsuyama, T. (2008). Speaker Detection Using the Timing Structure of Lip Motion and Sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

[Huang and Ahuja, 2012] Huang, J.-B. and Ahuja, N. (2012). Saliency Detection via Divergence Analysis: A Unified Perspective. In *Proceedings of the International Conference on Pattern Recogntion*, pages 2748–2751.

[Ishikawa et al., 2012] Ishikawa, E., Yonetani, R., Kawashima, H., Hirayama, T., and Matsuyama, T. (2012). Semantic Interpretation of Eye Movements using Designed Structures of Displayed Contents. In *Proceedings of the Workshop on Eye Gaze in Intelligent Human Machine Interaction*.

[Ishikawa et al., 2004] Ishikawa, T., Baker, S., Matthews, I., and Kanade, T. (2004). Passive Driver Gaze Tracking with Active Appearance Models. Technical report, Robotics Institute.

[Itti and Baldi, 2009] Itti, L. and Baldi, P. (2009). Bayesian Surprise Attracts Human Attention. *Vision Research*, 49(10):1295–1306.

[Itti et al., 2003] Itti, L., Dhavale, N., and Pighin, F. (2003). Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. In *Proceedings of the SPIE Annual International Symposium on Optical Science and Technology*, pages 64–78.

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

[Jacob and Karn, 2003] Jacob, R. J. K. and Karn, K. S. (2003). Commentary on Section 4. Eye Tracking in Human-computer Interaction and Usability Research: Ready to Deliver the Promises. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 573–605. Elsevier Science.

[Jayagopi and Gatica-Perez, 2010] Jayagopi, D. B. and Gatica-Perez, D. (2010). Mining Group Nonverbal Conversational Patterns Using Probabilistic Topic Models. *IEEE Transactions on Multimedia*, 12(8):790–802.

[Jayagopi et al., 2012] Jayagopi, D. B., Sanchez-Cortes, D., Otsuka, K., Yamato, J., and Gatica-Perez, D. (2012). Linking Speaking and Looking Behavior Patterns

with Group Composition, Perception, and Performance. In *Proceedings of the ACM International Conference on Multimodal interaction*, pages 433–440.

[Joiner and Shelhamer, 2006] Joiner, W. M. and Shelhamer, M. (2006). Pursuit and Saccadic Tracking Exhibit a Similar Dependence on Movement Preparation Time. *Experimental Brain Research*, 173(4):572–586.

[Judd et al., 2009] Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to Predict Where Humans Look. In *Proceedings of the IEEE International Conference on Computer Vision*.

[Kahneman, 1973] Kahneman, D. (1973). *Attention and Effort*. Prentice Hall.

[Kass et al., 1988] Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active Contour Models. *International Journal of Computer Vision*, 1(4):321–331.

[Keogh et al., 2005] Keogh, E., Lin, J., and Fu, A. (2005). Hot sax: efficiently finding the most unusual time series subsequence. In *Proceedings of the IEEE International Conference on Data Mining*.

[Keogh et al., 2003] Keogh, E., Lin, J., and Truppel, W. (2003). Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research. In *Proceedings of the IEEE International Conference on Data Mining*, pages 115–122.

[Kidron et al., 2007] Kidron, E., Schechner, Y., and Elad, M. (2007). Cross-modal Localization via Sparsity. *IEEE Transactions on Signal Processing*, 55(4):1390 – 1404.

[Kienzle et al., 2009] Kienzle, W., Franz, M. O., Schölkoph, B., and Wichmann, F. A. (2009). Center-surround Patterns Emerge as Optimal Predictors for Human Saccade Targets. *Journal of Vision*, 9(5):1–15.

[Kimura et al., 2013] Kimura, A., Yonetani, R., and Hirayama, T. (2013). Computational Models of Human Visual Attention and Their Implementations: A Survey. *IEICE Transactions on Information and Systems*, E96-D(3):562–578.

[Koch and Ullman, 1985] Koch, C. and Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4:219–227.

[Kowler, 2011] Kowler, E. (2011). Eye Movements: The Past 25 Years. *Vision Research*, 51(13):1457–83.

[Laptev, 2005] Laptev, I. (2005). On Space-time Interest Points. *International Journal of Computer Vision*, 64(2-3):107–123.

[Lee and Seung, 1999] Lee, D. and Seung, H. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401(6755):788–791.

[Lee and Seung, 2001] Lee, D. and Seung, H. (2001). Algorithms for Non-negative Matrix Factorization. In *Proceedings of the Conference on Neural Information Processing Systems*.

[Leykin and Tuceryan, 2004] Leykin, A. and Tuceryan, M. (2004). Automatic Determination of Text Readability over Textured Backgrounds for Augmented Reality Systems. In *Proceedings of the IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 224–230.

[Li et al., 2004] Li, L., Huang, W., Gu, I. Y., and Tian, Q. (2004). Statistical Modeling of Complex Backgrounds for Foreground Object Detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472.

[Li and Ngom, 2013] Li, Y. and Ngom, A. (2013). The Non-negative Matrix Factorization Toolbox for Biological Data Mining. *BMC Source Code for Biology and Medicine*, 8(1):10.

[Li et al., 2002] Li, Y., Wang, T., and Shum, H.-Y. (2002). Motion Texture: A Two-level Statistical Model for Character Motion Synthesis. *ACM Transactions on Graphics*, 21(3):465–472.

[Li et al., 2011] Li, Z., Yap, K.-H., and Chen, X.-M. (2011). Beyond Bag of Words: Combining Generative and Discriminative Models for Natural Scene Categorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 965–968.

[Liu and Sato, 2009] Liu, Y. and Sato, Y. (2009). Visual Localization of Non-Stationary Sound Sources. In *Proceedings of the ACM International Conference on Multimedia*, pages 513–516.

[Lowe, 1999] Lowe, D. G. (1999). Object Recognition from Local Scale-invariant Features. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157.

[Ma et al., 2005] Ma, Y.-F., Hua, X., Lu, L., and Zhang, H.-J. (2005). A Generic Framework of User Attention Model and Its Application in Video Summarization. *IEEE Transactions on Multimedia*, 7(5):907–919.

[Mahadevan and Vasconcelos, 2010] Mahadevan, V. and Vasconcelos, N. (2010). Spatiotemporal Saliency in Dynamic Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):171–177.

[Mai et al., 2013] Mai, L., Niu, Y., and Liu, F. (2013). Saliency Aggregation: A Data-Driven Approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Marat et al., 2009] Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., and Guérin-Dugué, A. (2009). Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos. *International Journal of Computer Vision*, 82(3):231–243.

[Mathot and Theeuwes, 2010] Mathot, S. and Theeuwes, J. (2010). Evidence for the Predictive Remapping of Visual Attention. *Experimental Brain Research*, 200(1):117–122.

[Morimoto and Mimica, 2005] Morimoto, C. H. and Mimica, M. R. M. (2005). Eye Gaze Tracking Techniques for Interactive Applications. *Computer Vision and Image Understanding*, 98(1):4–24.

[Munn et al., 2008] Munn, S. M., Stefano, L., and Pelz, J. B. (2008). Fixation-identification in Dynamic Scenes: Comparing an Automated Algorithm to Manual Coding. In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization*, pages 33–42.

[Nakano and Ishii, 2010] Nakano, Y. and Ishii, R. (2010). Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 139–148.

[Navalpakkam and Itti, 2005] Navalpakkam, V. and Itti, L. (2005). Modeling the Influence of Task on Attention. *Vision Research*, 45(2):205 – 231.

[Navalpakkam and Itti, 2007] Navalpakkam, V. and Itti, L. (2007). Search Goal Tunes Visual Features Optimally. *Neuron*, 53(4):605–617.

[Neisser and Beller, 1965] Neisser, U. and Beller, H. (1965). Searching through Word Lists. *British Journal of Psychology*, 56:349–358.

[North et al., 2000] North, B., Blake, A., Isard, M., and Rittscher, J. (2000). Learning and Classification of Complex Dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1016–1034.

[Orlosky et al., 2013] Orlosky, J., Kiyokawa, K., and Takemura, H. (2013). Dynamic Text Management for See-through Wearable and Heads-up Display Systems. In *Proceedings of the ACM International Conference on Intelligent User Interfaces*, pages 363–370.

[Ostendorf et al., 1996] Ostendorf, M., Digalakis, V., and Kimball, O. (1996). From HMM's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378.

[Palmer, 1994] Palmer, J. (1994). Set-size Effects in Visual Search: The Effect of Attention Is Independent of the Stimulus for Simple Tasks. *Vision Research*, 34(13):1703–1721.

[Pang et al., 2008] Pang, D., Kimura, A., Takeuchi, T., Yamato, J., and Kashino, K. (2008). A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network. In *Proceedings of the IEEE International Conference on Multimedia & Expo*.

[Park et al., 2012] Park, H. S., Jain, E., and Sheikh, Y. (2012). 3D Gaze Concurrences From Head-mounted Cameras. In *Proceedings of the Advances in Neural Information Processing Systems*.

[Parkhurst et al., 2002] Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the Role of Salience in the Allocation of Overt Visual Attention. *Vision Research*, 42(1):107–123.

[Pavlovic et al., 2000] Pavlovic, V., Garg, A., Rehg, J., and Huang, T. (2000). Multimodal Speaker Detection Using Error Feedback Dynamic Bayesian Networks. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 34–41.

[Peters and Itti, 2007] Peters, R. and Itti, L. (2007). Beyond Bottom-up: Incorporating Task-dependent Influences into a Computational Model of Spatial At-

tention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Poppe, 2010] Poppe, R. (2010). A Survey on Vision-based Human Action Recognition. *Image and Vision Computing*, 28(6):976 – 990.

[Rashbass, 1961] Rashbass, C. (1961). The Relationship between Saccadic and Smooth Tracking Eye Movements. *The Journal of Physiology*, 159:326–338.

[Ratanamahatana et al., 2005] Ratanamahatana, C., Lin, J., Gunopulos, D., and Keogh, E. (2005). Mining Time Series Data. In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, chapter 1. Kluwer Academic Publishers.

[Rath and Makur, 1999] Rath, G. and Makur, A. (1999). Iterative Least Squares and Compression based Estimations for a Four-parameter Linear Global Motion Model and Global Motion Compensation. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(7):1075–1099.

[Ravichandran et al., 2009] Ravichandran, A., Chaudhry, R., and Vidal, R. (2009). View-invariant Dynamic Texture Recognition Using a Bag of Dynamical Systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1651–1657.

[Rayner, 1975] Rayner, K. (1975). Parafoveal Identification during a Fixation in Reading. *Acta Psychologica*, 39(4):271 – 281.

[Riche et al., 2012] Riche, N., Mancas, M., and Culibrk, D. (2012). Dynamic Saliency Models and Human Attention: A Comparative Study on Videos. In *Proceedings of the Asian Conference on Computer Vision*.

[Sadeghi and Farhadi, 2011] Sadeghi, M. and Farhadi, A. (2011). Recognition Using Visual Phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1745–1752.

[Schauerte and Stiefelhagen, 2013] Schauerte, B. and Stiefelhagen, R. (2013). ′Wow!′ Bayesian Surprise for Salient Acoustic Event Detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

[Seo and Milanfar, 2009] Seo, H. J. and Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):1–27.

[Sharma et al., 2012] Sharma, G., Jurie, F., and Schmid, C. (2012). Discriminative Spatial Saliency for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

[Shimonishi et al., 2013] Shimonishi, K., Kawashima, H., Yonetani, R., Ishikawa, E., and Matsuyama, T. (2013). Learning Aspects of Interests from Gaze. In *Proceedings of the Workshop on Eye Gaze in Intelligent Human Machine Interaction*.

[Simola et al., 2008] Simola, J., Salojärvi, J., and Kojo, I. (2008). Using Hidden Markov Model to Uncover Processing States from Eye Movements in Information Search Tasks. *Cognitive Systems Research*, 9(4):237–251.

[Simonin et al., 2005] Simonin, J., Kieffer, S., and Carbonell, N. (2005). Effects of Display Layout on Gaze Activity During Visual Search. In *Proceedings of the Human-Computer Interaction*, volume 3585, pages 1054–1057.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2.

[Smaragdis and Brown, 2003] Smaragdis, P. and Brown, J. (2003). Non-negative Matrix Factorization for Polyphonic Music Transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[Subramanian et al., 2011] Subramanian, R., Yanulevskaya, V., and Sebe, N. (2011). Can Computers Learn from Humans to See Better?: Inferring Scene Semantics from Viewers' Eye Movements. In *Proceedings of the ACM International Conference on Multimedia*, pages 33–42.

[Sun et al., 2007] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. (2007). GraphScope: Parameter-Free Mining of Large Time-Evolving Graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 687–696.

[Sun et al., 2012] Sun, X., Yao, H., and Ji, R. (2012). What Are We Looking for: Towards Statistical Modeling of Saccadic Eye Movements and Visual Saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1552–1559.

[Torralba et al., 2006] Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual Guidance of Eye Movements and Attention in Real-World Scenes: the Role of Global Features in Object Search. *Psychological Review*, 113(4):766–786.

[Treisman and Gelade, 1980] Treisman, A. and Gelade, G. (1980). A feature-integration Theory of Attention. *Cognitive Psychology*, 12:97–136.

[Tseng et al., 2013] Tseng, P.-H., Cameron, I. G., Pari, G., Reynolds, J. N., Munoz, D. P., and Itti, L. (2013). High-throughput Classification of Clinical Populations from Natural Viewing Eye Movements. *Journal of Neurology*, 260(1):275–284.

[Tsotsos, 2011] Tsotsos, J. K. (2011). *A Computational Perspective on Visual Attention*. MIT Press.

[Tsotsos et al., 1995] Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling Visual Attention via Selective Tuning. *Artificial Intelligence*, 78(1-2):507–545.

[Uřičář et al., 2012] Uřičář, M., Franc, V., and Hlávac, V. (2012). Detector of Facial Landmarks Learned by the Structured Output SVM. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

[Vijayakumar et al., 2001] Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. (2001). Overt Visual Attention for a Humanoid Robot. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 4, pages 2332–2337.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Vision Society of Japan, 2000] Vision Society of Japan., editor (2000). *Handbook of Visual Information Processing (in Japanese)*, chapter 9. Eye Movement. Asakura Publishing Co., Ltd.

[Wang et al., 2012] Wang, J.-C., Yang, Y.-H., Wang, H.-M., and Jeng, S.-K. (2012). The Acoustic Emotion Gaussians Model for Emotion-based Music Annotation and Retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 89–98.

[Wang et al., 2005] Wang, J.-G., Sung, E., and Venkateswarlu, R. (2005). Estimating the Eye Gaze from One Eye. *Computer Vision and Image Understanding*, 98(1):83–103.

[Witkin, 1983] Witkin, A. P. (1983). Scale-space Filtering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1019–1022.

[Wolfe, 1998] Wolfe, J. M. (1998). What Can 1 Million Trials Tell Us About Visual Search? *Psychological Science*, 9(1):33–39.

[Wright and Ward, 2008] Wright, R. D. and Ward, L. M. (2008). 6. Eye Movements and Attention Shifts. In *Orienting of Attention*, pages 121–152. Oxford University Press.

[Xu et al., 2003] Xu, W., Liu, X., and Gong, Y. (2003). Document Clustering Based on Non-negative Matrix Factorization. In *Proceedings of the ACM SIGIR Conference*.

[Yamada et al., 2010] Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., and Hiraki, K. (2010). Can Saliency Map Models Predict Human Egocentric Visual Attention? In *Proceedings of the Asian Conference on Computer Vision*, pages 420–429.

[Yang et al., 2011] Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting Communities and Their Evolutions in Dynamic Social Networks–a Bayesian Approach. *Machine Learning*, 82(2):157–189.

[Yarbus, 1967] Yarbus, A. (1967). Eye Movements and Vision. *Plenum*.

[Yi and Faloutsos, 2000] Yi, B.-K. and Faloutsos, C. (2000). Fast Time Sequence Indexing for Arbitrary Lp Norms. In *Proceedings of the International Conference on Very Large Data Bases*, pages 385–394.

[Yilmaz et al., 2006] Yilmaz, A., Javed, O., and Shah, M. (2006). Object Tracking: A Survey. *ACM Computer Surveys*, 38(4).

[Yonetani et al., 2010] Yonetani, R., Kawashima, H., Hirayama, T., and Matsuyama, T. (2010). Gaze Probing: Event-based Estimation of Objects Being Focused on. *The Transaction of Human Interface Society*, 12(3):125–135.

[Yonetani et al., 2013] Yonetani, R., Kawashima, H., and Matsuyama, T. (2013). Predicting Where We Look from Spatiotemporal Gaps. In *Proceedings of the International Conference on Multimodal Interaction*.

[Yoshitaka et al., 2007] Yoshitaka, A., Wakiyama, K., and Hirashima, T. (2007). Recommendation of Visual Information by Gaze-based Implicit Preference Acquisition. In *Proceedings of the Advances in Multimedia Modeling (MMM2007)*, pages 126–137.

[Yu et al., 2012] Yu, Z., Yu, Z., Zhou, X., Becker, C., and Nakamura, Y. (2012). Tree-Based Mining for Discovering Patterns of Human Interaction in Meetings. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):759–768.

[Zhan et al., 2008] Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A., and Xu, L.-Q. (2008). Crowd Analysis: A Survey. *Machine Vision and Applications*, 19(5-6):345–357.

[Zhang et al., 2013] Zhang, C., Liang, X., and Matsuyama, T. (2013). Mixed-Motion Segmentation Using Helmholtz Decomposition. *IPSJ Transactions on Computer Vision and Applications*, 5:55–59.

[Zhang et al., 2009] Zhang, L., Tong, M. H., and W, G. (2009). SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes. In *Proceedings of the Annual Cognitive Science Society Conference*.

[Zhu and Ji, 2005] Zhu, Z. and Ji, Q. (2005). Eye Gaze Tracking under Natural Head Movements. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1:918–923.

List of Publications

Journal papers

1. Ryo Yonetani, Hiroaki Kawashima and Takashi Matsuyama: "Learning Spatiotemporal Gaps between Where We Look and What We Focus on" IPSJ Transactions on Computer Vision and Applications, 5, pp. 75-79, 2013.

2. Ryo Yonetani, Hiroaki Kawashima, Takekazu Kato and Takashi Matsuyama: "Modeling Video Saliency Dynamics for Viewer State Estimation", The Transaction of The Institute of Electronics, Information and Communication Engineers, J96-D(8), pp.1675-1687, 2013 (Recommended paper; in Japanese).

3. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Mental Focus Analysis Using the Spatio-temporal Correlation between Visual Saliency and Eye Movements", Journal of Information Processing, 20(1), pp.267-276, 2012.

4. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Gaze Probing: Event-based Estimation of Objects Being Focused on", The Transaction of Human Interface Society, 12(3), pp. 125-135, 2010 (in Japanese).

International Conference (reviewed)

1. Ryo Yonetani, Hiroaki Kawashima and Takashi Matsuyama: "Predicting Where We Look from Spatiotemporal Gaps", International Conference on Multimodal Interaction (ICMI 2013), Sidney, Australia, Dec. 2013.

2. Ryo Yonetani: "Modeling Video Viewing Behaviors for Viewer State Estimation", ACM Multimedia Doctoral Symposium (ACMMM 2012 DS), Nara, Japan, Oct. 2012 (Award candidate in the Doctoral Symposium).

3. Ryo Yonetani, Hiroaki Kawashima and Takashi Matsuyama: "Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention", Eye Tracking Research & Applications (ETRA 2012), Santa Barbara, CA, USA, Mar. 2012.

4. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Gaze Probing: Event-Based Estimation of Objects Being Focused On", International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, Aug. 2010 (IBM Best Student Paper Award).

Presentation (in Japanese)

1. Ryo Yonetani, Hiroaki Kawashima and Takashi Matsuyama: "Gaze Point Prediction with Gap Structure Models", Technical Report of IEICE, PRMU, Sep. 2013.

2. Ryo Yonetani, Hiroaki Kawashima, Takekazu Kato and Takashi Matsuyama: "Modeling Video Saliency Dynamics for Viewer State Estimation", Meeting of Image Recognition and Understanding (MIRU2012), Aug. 2012 (Best Student Paper Award).

3. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Mental Focus Estimation Using the Spatiotemporal Correlation between Video Saliency and Eye Movements ", Technical Report of IPSJ, CVIM178-16, Sep. 2011.

4. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Dynamic Content Design for the Estimation of the Gazed Objects and Its Evaluation", National Convention of IPSJ, pp. 5-141-142, Mar. 2010.

5. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Gaze Probing: Event-based Estimation of Objects Being Focused on", Meeting of Image Recognition and Understanding (MIRU2009), pp.1713-1720, Jul. 2009.

6. Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama and Takashi Matsuyama: "Gazed Object Estimation Using the Timing Structure between Displayed Events and Eye Movements", Technical Report of IPSJ, CVIM 167-16, Jun. 2009.